

Ivan Dimov
Stefka Dimova
Natalia Kolkovska (Eds.)

LNCS 6046

Numerical Methods and Applications

7th International Conference, NMA 2010
Borovets, Bulgaria, August 2010
Revised Papers

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Ivan Dimov Stefka Dimova
Natalia Kolkovska (Eds.)

Numerical Methods and Applications

7th International Conference, NMA 2010
Borovets, Bulgaria, August 20-24, 2010
Revised Papers



Springer

Volume Editors

Ivan Dimov
Bulgarian Academy of Sciences
Institute of Computer and Communication Technologies
Acad. G. Bonchev 25 A, 1113 Sofia, Bulgaria
E-mail: ivdimov@bas.bg

Stefka Dimova
University of Sofia "St. Kliment Ohridski"
Faculty of Mathematics and Informatics
Department Numerical Methods and Algorithms
Blvd. James Bourchier 5, 1164 Sofia, Bulgaria
E-mail: dimova@fmi.uni-sofia.bg

Natalia Kolkovska
Bulgarian Academy of Sciences
Institute of Mathematics and Informatics
Acad. Bonchev St., Bl. 8, 1113 Sofia, Bulgaria
E-mail: natali@math.bas.bg

ISSN 0302-9743
ISBN 978-3-642-18465-9
DOI 10.1007/978-3-642-18466-6
Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349
e-ISBN 978-3-642-18466-6

Library of Congress Control Number: 2010942928

CR Subject Classification (1998): G.1, F.2.1, G.4, I.6, J.2, J.6

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The international conference Numerical Methods and Applications is a traditional forum for scientists from all over the world providing an opportunity to share ideas and establish fruitful scientific cooperation. The aim of the conference is to bring together leading international scientists of the numerical and applied mathematics community and to attract original research papers of very high quality.

The papers in this volume were presented at the seventh edition of the International Conference on Numerical Methods and Applications (ICNM&A 2010) held in Borovets, Bulgaria, August 20–24, 2010. The conference was organized by the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences in cooperation with SIAM. The Faculty of Mathematics and Informatics of St. Kliment Ohridski University of Sofia and the Institute of Computer and Communication Technologies, Bulgarian Academy of Sciences were co-organizers of this traditional scientific meeting. Over 100 participants from 22 countries attended the conference. Ninety-four talks, including ten invited and keynote talks, were presented. This volume contains 60 papers submitted by authors from 16 countries.

During ICNM&A 2010 a wide range of problems concerning recent theoretical achievements in numerical methods and their applications in mathematical modeling were discussed. Specific topics of interest were the following: *Numerical methods for differential and integral equations; approximation techniques in numerical analysis; numerical linear algebra; hierarchical and domain decomposition methods; parallel algorithms; Monte Carlo methods; computational mechanics; computational physics, chemistry and biology; engineering applications*. Five special sessions were organized: *Monte Carlo and Quasi-Monte Carlo Methods; Environmental Modeling; Grid Computing and Applications; Metaheuristics for Optimisation Problems; Modeling and Simulation of Electrochemical Processes*. The ICNM&A 2010 talks were delivered by researchers representing some of the strongest research teams in the field of numerical methods and their application for solving a wide range of practical problems.

The success of the conference and the present volume are due to the joint efforts of many colleagues from various institutions and organizations. We express our deep gratitude to all the members of the Scientific Committee for their valuable contribution to forming the scientific spirit of the conference, as well as for their help in reviewing the submitted papers. We are also grateful to the staff involved in the local organization.

We hope that this meeting among scientists who develop and study numerical methods, on one hand, and researchers who use them for solving real-life problems, on the other, has broadened their horizons and contributed to their mutual enrichment.

December 2010

Ivan Dimov
Stefka Dimova
Natalia Kolkovska

Organization

International Scientific Committee

A. Andreev (Bulgaria)	J. Guermond (USA)	S. Radev (Bulgaria)
E. Atanassov (Bulgaria)	R. Herbin (France)	P. Ribeiro (Portugal)
R. Blaheta (Czech Republic)	O. Iliev (Germany)	K. Sabelfeld (Russia)
T. Boyadjiev (Bulgaria)	B. Jovanovic (Serbia)	J. Schoeberl (Germany)
J. Buša (Slovakia)	S. Korotov (Finland)	S. Selberherr (Austria)
R. Ciegis (Lithuania)	J. Kraus (Austria)	Bl. Sendov (Bulgaria)
P. D'Ambra (Italy)	N. Krejic (Serbia)	K. Semerdzhiev (Bulgaria)
I. Dimov (Bulgaria)	R. Lazarov (USA)	S. Slavchev (Bulgaria)
S. Dimova (Bulgaria)	I. Lirkov (Bulgaria)	M. Todorov (Bulgaria)
I. Farago (Hungary)	S. Margenov (Bulgaria)	V. Thomee (Sweden)
M. Feistauer (Czech Republic)	P. Marinov (Bulgaria)	P. Vabishchevich (Russia)
S. Fidanova (Bulgaria)	S. Markov (Bulgaria)	I. Yotov (USA)
K. Georgiev (Bulgaria)	P. Matus (Belarus)	L. Zikatanov (USA)
A. Goolin (Russia)	P. Minev (Canada)	
S. Gocheva-Ilieva (Bulgaria)	M. Nedjalkov (Bulgaria)	
	J. Pedroso (Portugal)	
	K. Penev (UK)	
	B. Popov (USA)	

Organizing Committee

Chairperson: N. Kolkovska

I. Bazhlekov	M. Dimova	S. Stoilova
T. Chernogorova	I. Georgiev	D. Vasileva
I. Christov		

Table of Contents

Invited Papers

Space-Time Discontinuous Galerkin Finite Element Method for Convection-Diffusion Problems and Compressible Flow	1
<i>Miloslav Feistauer and Jan Česenek</i>	
Stochastic Algorithms in Linear Algebra - beyond the Markov Chains and von Neumann - Ulam Scheme.....	14
<i>Karl Sabelfeld</i>	
SM Stability for Time-Dependent Problems	29
<i>Petr N. Vabishchevich</i>	

Monte Carlo and Quasi-Monte Carlo Methods

Advanced Monte Carlo Techniques in the Simulation of CMOS Devices and Circuits.....	41
<i>Asen Asenov</i>	
Monte Carlo Method for Numerical Integration Based on Sobol's Sequences	50
<i>Ivan Dimov and Rayna Georgieva</i>	
Using Monte-Carlo Simulation for Risk Assessment: Application to Occupational Exposure during Remediation Works	60
<i>M.L. Dinis and A. Fiúza</i>	
The b -adic Diaphony as a Tool to Study Pseudo-randomness of Nets....	68
<i>Ivan Lirkov and Stanislava Stoilova</i>	
Scatter Estimation for PET Reconstruction	77
<i>Milan Magdics, Laszlo Szirmay-Kalos, Balazs Tóth, Ádam Csenedesi, and Anton Penzov</i>	
Modeling of the SET and RESET Process in Bipolar Resistive Oxide-Based Memory Using Monte Carlo Simulations	87
<i>Alexander Makarov, Viktor Sverdlov, and Siegfried Selberherr</i>	
Stochastic Algorithm for Solving the Wigner-Boltzmann Correction Equation.....	95
<i>M. Nedjalkov, S. Selberherr, and I. Dimov</i>	

Modeling Thermal Effects in Fully-Depleted SOI Devices with Arbitrary Crystallographic Orientation	103
<i>K. Raleva, D. Vasilevska, and S.M. Goodnick</i>	
Particle Monte Carlo Algorithms with Small Number of Particles in Grid Cells	110
<i>Stefan K. Stefanov</i>	
Is Self-Heating Important in Nanowire FETs?	118
<i>D. Vasilevska, A. Hossain, K. Raleva, and S.M. Goodnick</i>	
Environmental Modeling	
Mixed-Hybrid Formulation of Multidimensional Fracture Flow	125
<i>Jan Březina and Milan Hokr</i>	
WRF-Fire Applied in Bulgaria	133
<i>Nina Dobrinkova, Georgi Jordanov, and Jan Mandel</i>	
Bulgarian Operative System for Chemical Weather Forecast	141
<i>Iglika Etropolska, Maria Prodanova, Dimiter Syrakov, Kostadin Ganev, Nikolai Miloshev, and Kiril Slavov</i>	
Atmospheric Composition Studies for the Balkan Region	150
<i>Georgi Gadzhev, Georgi Jordanov, Kostadin Ganev, Maria Prodanova, Dimiter Syrakov, and Nikolai Miloshev</i>	
Specialized Sparse Matrices Solver in the Chemical Part of an Environmental Model	158
<i>Krassimir Georgiev and Zahari Zlatev</i>	
A Numerical Investigation for the Optimal Contaminant Inlet Positions in Horizontal Subsurface Flow Wetlands	167
<i>Konstantinos Liolios, Vassilios Tsihrintzis, and Stefan Radev</i>	
Using Satellite Observations for Air Quality Assessment with an Inverse Model System	174
<i>Achim Strunk, Hendrik Elbern, and Adolf Ebel</i>	
Distributed Software System for Data Evaluation and Numerical Simulations of Atmospheric Processes	182
<i>Atanas T. Terziyski and Nikolay T. Kochev</i>	
Advanced Numerical Tools Applied to Geo-environmental Engineering - Soils Contaminated by Petroleum Hydrocarbons, a Case Study	190
<i>Maria Cristina Vila, J.M. Soeiro de Carvalho, and António Fiúza</i>	

Richardson Extrapolated Numerical Methods for Treatment of One-Dimensional Advection Equations	198
<i>Zahari Zlatev, Ivan Dimov, István Faragó, Krassimir Georgiev, Ágnes Havasi, and Tzvetan Ostromsky</i>	

Grid Computing and Applications

Programming Problems with a Large Number of Objective Functions . . .	207
<i>Cornel Resteanu and Romica Trandafir</i>	
First Results of SEE-GRID-SCI Application CCIAQ	215
<i>Dimiter Syrakov, Valery Spiridonov, Kostadin Ganev, Maria Prodanova, Andrey Bogachev, Nikolai Miloshev, and Kiril Slavov</i>	

Metaheuristics for Optimization Problems

Genetic Algorithms Based Parameter Identification of Yeast Fed-Batch Cultivation	224
<i>Maria Angelova, Stoyan Tzonkov, and Tania Pencheva</i>	
Intuitionistic Fuzzy Interpretations of Conway's Game of Life	232
<i>Lilija Atanassova and Krassimir Atanassov</i>	
Ant Colony Optimization Approach to Tokens' Movement within Generalized Nets	240
<i>Vassia Atanassova and Krassimir Atanassov</i>	
Start Strategies of ACO Applied on Subset Problems	248
<i>Stefka Fidanova, Krassimir Atanassov, and Pencho Marinov</i>	
Sensitivity Analysis of ACO Start Strategies for Subset Problems	256
<i>Stefka Fidanova, Pencho Marinov, and Krassimir Atanassov</i>	
A Highly-Parallel TSP Solver for a GPU Computing Platform	264
<i>Noriyuki Fujimoto and Shigeyoshi Tsutsui</i>	
Metaheuristics for the Asymmetric Hamiltonian Path Problem	272
<i>João Pedro Pedroso</i>	
Adaptive Intelligence Applied to Numerical Optimisation	280
<i>Kalin Penev and Anton Ruzhekov</i>	
Fed-Batch Cultivation Control Based on Genetic Algorithm PID Controller Tuning	289
<i>Olympia Roeva and Tsonyo Slavov</i>	
Perspectives of Selfish Behaviour in Mobile Ad Hoc Networks	297
<i>Marcin Seredynski and Pascal Bouvry</i>	

A Comparison of Metaheuristics for the Problem of Solving Parametric Interval Linear Systems	305
<i>Iwona Skalna and Jerzy Duda</i>	

Parametric Approximation of Functions Using Genetic Algorithms: An Example with a Logistic Curve	313
<i>Fernando Torrecilla-Pinero, Jesús A. Torrecilla-Pinero, Juan A. Gómez-Pulido, Miguel A. Vega-Rodríguez, and Juan M. Sánchez-Pérez</i>	

Population-Based Metaheuristics for Tasks Scheduling in Heterogeneous Distributed Systems	321
<i>Flavia Zamfirache, Marc Frîncu, and Daniela Zaharie</i>	

Modeling and Simulation of Electrochemical Processes

Modeling of Species and Charge Transport in Li-Ion Batteries Based on Non-equilibrium Thermodynamics	329
<i>Arnulf Latz, Jochen Zausch, and Oleg Iliev</i>	

Finite Volume Discretization of Equations Describing Nonlinear Diffusion in Li-Ion Batteries	338
<i>P. Popov, Y. Vutov, S. Margenov, and O. Iliev</i>	

Contributed Papers

Numerical Study of Magnetic Flux in the LJJ Model with Double Sine-Gordon Equation	347
<i>P.Kh. Atanasova, T.L. Boyadjiev, E.V. Zemlyanaya, and Yu.M. Shukrinov</i>	

A Simple Preconditioner for the SIPG Discretization of Linear Elasticity Equations	353
<i>B. Ayuso, I. Georgiev, J. Kraus, and L. Zikatanov</i>	

Merger Bound States in $0 - \pi$ Josephson Structures	361
<i>Todor L. Boyadjiev and Hristo T. Melemov</i>	

Some Error Estimates for the Discretization of Parabolic Equations on General Multidimensional Nonconforming Spatial Meshes	369
<i>Abadallah Bradji and Jürgen Fuhrmann</i>	

Finite-Volume Difference Scheme for the Black-Scholes Equation in Stochastic Volatility Models	377
<i>Tatiana Chernogorova and Radoslav Valkov</i>	

On the Numerical Simulation of Unsteady Solutions for the 2D Boussinesq Paradigm Equation	386
<i>Christo I. Christov, Natalia Kolkovska, and Daniela Vasileva</i>	
Numerical Investigation of Spiral Structure Solutions of a Nonlinear Elliptic Problem	395
<i>Milena Dimova and Stefka Dimova</i>	
Bidirectional Beam Propagation Method Applied for Lasers with Multilayer Active Medium	404
<i>N.N. Elkin, A.P. Napartovich, and D.V. Vysotsky</i>	
Analysis of the CBS Constant for Quadratic Finite Elements	412
<i>Ivan Georgiev, Maria Lymbery, and Svetozar Margenov</i>	
Sensitivity of Results of the Water Flow Problem in a Discrete Fracture Network with Large Coefficient Differences	420
<i>Milan Hokr, Jiří Kopal, Jan Březina, and Petr Rálek</i>	
Fluxon Dynamics in Stacked Josephson Junctions	428
<i>Ivan Hristov and Stefka Dimova</i>	
Global Convergence Properties of the SOR-Weierstrass Method	437
<i>Vladimir Hristov, Nikolay Kyurkchiev, and Anton Iliev</i>	
Numerical Solution of a Nonlinear Evolution Equation for the Risk Preference	445
<i>Naoyuki Ishimura, Miglena N. Koleva, and Lubin G. Vulkov</i>	
A Numerical Approach for the American Call Option Pricing Model	453
<i>Juri D. Kandilarov and Radoslav L. Valkov</i>	
A Numerical Study of a Parabolic Monge-Ampère Equation in Mathematical Finance	461
<i>Miglena N. Koleva and Lubin G. Vulkov</i>	
Convergence of Finite Difference Schemes for a Multidimensional Boussinesq Equation	469
<i>Natalia T. Kolkovska</i>	
A Numerical Approach for Obtaining Fragility Curves in Seismic Structural Mechanics: A Bridge Case of Egnatia Motorway in Northern Greece	477
<i>Asterios Liolios, Panagiotis Panetsos, Angelos Liolios, George Hatzigeorgiou, and Stefan Radev</i>	
An Efficient Numerical Method for a System of Singularly Perturbed Semilinear Reaction-Diffusion Equations	486
<i>S. Chandra Sekhara Rao and Sunil Kumar</i>	

A Comparison of Methods for Solving Parametric Interval Linear Systems with General Dependencies	494
<i>Iwona Skalna</i>	
Numerical Investigation of the Upper Bounds on the Convective Heat Transport in a Heated from below Rotating Fluid Layer	502
<i>Nikolay Vitanov</i>	
Author Index	511

Space-Time Discontinuous Galerkin Finite Element Method for Convection-Diffusion Problems and Compressible Flow

Miloslav Feistauer and Jan Česenek

Charles University Prague, Faculty of Mathematics and Physics,
Sokolovská 83, 186 75 Praha 8, Czech Republic
feist@karlin.mff.cuni.cz, jan.cessa@seznam.cz

Abstract. This paper is concerned with the numerical solution of non-stationary, nonlinear, convection-diffusion problems by the space-time discontinuous Galerkin finite element method (DGFEM) and applications to compressible flow. The first part is devoted to theoretical analysis of error estimates of the method. In the second part, this technique is applied to the numerical solution of compressible flow in time-dependent domains and the simulation of flow induced airfoil vibrations.

Keywords: nonlinear nonstationary convection-diffusion problems, space-time discontinuous Galerkin discretization, error estimates, numerical solution of compressible flow in time-dependent domains, ALE method, airfoil vibrations.

1 Introduction

During the last decade the discontinuous Galerkin finite element method, using piecewise polynomial discontinuous approximations (cf., e.g. [2]), appeared as an efficient tool for the space discretization of a number of problems described by partial differential equations. The numerical simulation of strongly nonstationary transient problems requires the application of numerical schemes of high order of accuracy both in space and in time. From this point of view, it appears suitable to use the discontinuous Galerkin discretization with respect to space as well as time. The discontinuous Galerkin time discretization was introduced and analyzed, e.g. in [9] for the solution of ordinary differential equations. In [10] and references therein, the solution of linear parabolic problems is carried out with the aid of conforming finite elements in space combined with the DG time discretization. In [5], the space-time DGFEM was analyzed for a linear nonstationary convection-diffusion-reaction problem. The papers [6] and [7] are devoted to the analysis of a nonstationary convection-diffusion problem with a nonlinear convection and linear diffusion. In the present paper we are concerned with the space-time discontinuous Galerkin discretization applied to the numerical solution of a nonstationary convection-diffusion problem with a nonlinear convection as well as diffusion. In the second part of the paper we apply this

method to the simulation of compressible flow in time-dependent domains and flow induced airfoil vibrations. For simplicity we shall consider problems with two space dimensions.

We consider the following initial-boundary value problem. Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain and $T > 0$. We want to find $u : Q_T = \Omega \times (0, T) \rightarrow \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^2 \frac{\partial f_s(u)}{\partial x_s} - \operatorname{div}(\beta(u) \nabla u) = g \quad \text{in } Q_T, \quad (1)$$

$$u|_{\partial\Omega \times (0, T)} = u_D, \quad (2)$$

$$u(x, 0) = u^0(x), \quad x \in \Omega. \quad (3)$$

We assume that g, u_D, u^0, f_s are given functions and $f_s \in C^1(\mathbb{R})$, $|f'_s| \leq C$, $s = 1, 2$. Moreover, let

$$\beta : \mathbb{R} \rightarrow [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \quad (4)$$

$$|\beta(u_1) - \beta(u_2)| \leq L|u_1 - u_2| \quad \forall u_1, u_2 \in \mathbb{R}. \quad (5)$$

In the derivation and analysis of the discrete problem we assume that the exact solution is regular in the following sense:

$$u \in L^2(0, T; H^2(\Omega)), \quad \frac{\partial u}{\partial t} \in L^2(0, T; H^1(\Omega)), \quad (6)$$

$$\|\nabla u(t)\|_{L^\infty(\Omega)} \leq C_R \quad \text{for a.e. } t \in (0, T). \quad (7)$$

2 Space-Time Discretization

In the time interval $[0, T]$ we shall construct a partition $0 = t_0 < \dots < t_M = T$ and denote $I_m = (t_{m-1}, t_m)$, $\tau_m = t_m - t_{m-1}$, $\tau = \max_{m=1, \dots, M} \tau_m$. For each I_m we consider a partition $\mathcal{T}_{h,m}$ of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed triangles with mutually disjoint interiors. The partitions $\mathcal{T}_{h,m}$ are in general different for different m .

By $\mathcal{F}_{h,m}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{h,m}$. Further, we denote the set of all inner faces by $\mathcal{F}_{h,m}^I$ and the set of all boundary faces by $\mathcal{F}_{h,m}^B$. Each $\Gamma \in \mathcal{F}_{h,m}$ will be associated with a unit normal vector $\mathbf{n}_\Gamma$, which has the same orientation as the outer normal to $\partial\Omega$ for $\Gamma \in \mathcal{F}_{h,m}^B$. We set $h_K = \operatorname{diam}(K)$ for $K \in \mathcal{T}_{h,m}$, $h_m = \max_{K \in \mathcal{T}_{h,m}} h_K$, $h = \max_{m=1, \dots, M} h_m$. By ρ_K we denote the radius of the largest circle inscribed into K .

For a function φ defined in $\bigcup_{m=1}^M I_m$ we put $\varphi_m^\pm = \varphi(t_m \pm) = \lim_{t \rightarrow t_m \pm} \varphi(t)$ and $\{\varphi\}_m = \varphi(t_m+) - \varphi(t_m-)$.

Over a triangulation $\mathcal{T}_{h,m}$ we define the *broken Sobolev spaces* $H^k(\Omega, \mathcal{T}_{h,m}) = \{v; v|_K \in H^k(K) \forall K \in \mathcal{T}_{h,m}\}$. For each face $\Gamma \in \mathcal{F}_{h,m}^I$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_{h,m}$ such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to $\partial K_\Gamma^{(L)}$ and the inner normal to $\partial K_\Gamma^{(R)}$. If

$\Gamma \in \mathcal{F}_{h,m}^B$, then $K_\Gamma^{(L)}$ will denote the element adjacent to Γ . For $v \in H^1(\Omega, \mathcal{T}_{h,m})$ and $\Gamma \in \mathcal{F}_{h,m}$ we use the notation $v_\Gamma^{(L)}$ for the trace of $v|_{K_\Gamma^{(L)}}$ on Γ . If $\Gamma \in \mathcal{F}_{h,m}^I$, then we set $v_\Gamma^{(R)}$ = the trace of $v|_{K_\Gamma^{(R)}}$ on Γ , $\langle v \rangle_\Gamma = \frac{1}{2} \left(v_\Gamma^{(L)} + v_\Gamma^{(R)} \right)$, $[v]_\Gamma = v_\Gamma^{(L)} - v_\Gamma^{(R)}$.

Let $C_W > 0$ be a fixed constant. We set

$$h(\Gamma) = \frac{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}{2C_W} \quad \text{for } \Gamma \in \mathcal{F}_{h,m}^I, \quad h(\Gamma) = \frac{h_{K_\Gamma^{(L)}}}{C_W} \quad \text{for } \Gamma \in \mathcal{F}_{h,m}^B. \quad (8)$$

By $(\cdot, \cdot)$ we denote the scalar product in $L^2(\Omega)$ and by $\|\cdot\|$ we denote the norm in $L^2(\Omega)$. If $u, v, \varphi \in H^2(\Omega, \mathcal{T}_{h,m})$, we define the forms

$$a_{h,m}(v, u, \varphi) = \sum_{K \in \mathcal{T}_{h,m}} \int_K \beta(v) \nabla u \cdot \nabla \varphi \, dx \quad (9)$$

$$\begin{aligned} & - \sum_{\Gamma \in \mathcal{F}_{h,m}^I} \int_\Gamma (\langle \beta(v) \nabla u \rangle \cdot \mathbf{n}_\Gamma [\varphi] + \theta \langle \beta(v) \nabla \varphi \rangle \cdot \mathbf{n}_\Gamma [u]) \, dS \\ & - \sum_{\Gamma \in \mathcal{F}_{h,m}^B} \int_\Gamma (\beta(v) \nabla u \cdot \mathbf{n}_\Gamma \varphi + \theta \beta(v) \nabla \varphi \cdot \mathbf{n}_\Gamma u - \theta \beta(v) \nabla \varphi \cdot \mathbf{n}_\Gamma u_D) \, dS, \end{aligned}$$

$$J_{h,m}(u, \varphi) = \sum_{\Gamma \in \mathcal{F}_{h,m}^I} h(\Gamma)^{-1} \int_\Gamma [u] [\varphi] \, dS + \sum_{\Gamma \in \mathcal{F}_{h,m}^B} h(\Gamma)^{-1} \int_\Gamma u \varphi \, dS, \quad (10)$$

$$A_{h,m} = a_{h,m} + \beta_0 J_{h,m}, \quad (11)$$

$$b_{h,m}(u, \varphi) = - \sum_{K \in \mathcal{T}_{h,m}} \int_K \sum_{s=1}^2 f_s(u) \frac{\partial \varphi}{\partial x_s} \, dx \quad (12)$$

$$+ \sum_{\Gamma \in \mathcal{F}_{h,m}^I} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \mathbf{n}_\Gamma) [\varphi] \, dS + \sum_{\Gamma \in \mathcal{F}_{h,m}^B} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(L)}, \mathbf{n}_\Gamma) \varphi \, dS.$$

$$\ell_{h,m}(\varphi) = (g, \varphi) + \beta_0 \sum_{\Gamma \in \mathcal{F}_{h,m}^B} h(\Gamma)^{-1} \int_\Gamma u_D \varphi \, dS. \quad (13)$$

In (12), H is a numerical flux with the following properties.

(H1) $H(u, v, \mathbf{n})$ is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{\mathbf{n} \in \mathbb{R}^2; |\mathbf{n}| = 1\}$, and is Lipschitz-continuous with respect to u, v .

(H2) $H(u, v, \mathbf{n})$ is consistent: $H(u, u, \mathbf{n}) = \sum_{s=1}^2 f_s(u) n_s$, $u \in \mathbb{R}$, $\mathbf{n} = (n_1, n_2) \in B_1$.

(H3) $H(u, v, \mathbf{n})$ is conservative: $H(u, v, \mathbf{n}) = -H(v, u, -\mathbf{n})$, $u, v \in \mathbb{R}$, $\mathbf{n} \in B_1$.

In the above forms we take $\theta = -1$, $\theta = 0$ and $\theta = 1$ and obtain the non-symmetric (NIPG), incomplete (IIPG) and symmetric (SIPG) variants of the approximation of the diffusion terms, respectively.

In the space $H^1(\Omega, \mathcal{T}_{h,m})$, the following norm will be used:

$$\|\varphi\|_{DG,m} = \left(\sum_{K \in \mathcal{T}_{h,m}} |\varphi|_{H^1(K)}^2 + J_{h,m}(\varphi, \varphi) \right)^{1/2}. \quad (14)$$

Let $p, q \geq 1$ be integers. For each $m = 1, \dots, M$ we define the finite-dimensional space

$$S_{h,m}^p = \{ \varphi \in L^2(\Omega); \varphi|_K \in P^p(K) \ \forall K \in \mathcal{T}_{h,m} \}. \quad (15)$$

Here $P^p(K)$ denotes the space of all polynomials on K of degree $\leq p$. We denote by Π_m the $L^2(\Omega)$ -projection on $S_{h,m}^p$. The approximate solution will be sought in the space

$$S_{h,\tau}^{p,q} = \left\{ \varphi \in L^2(Q_T); \varphi|_{I_m} = \sum_{i=0}^q t^i \varphi_i \quad \text{with } \varphi_i \in S_{h,m}^p, \ m = 1, \dots, M \right\}. \quad (16)$$

In what follows we shall use the notation $U' = \partial U / \partial t$, $u' = \partial u / \partial t$.

Definition 1. We say that a function U is an approximate solution of problem (1) – (3), if $U \in S_{h,\tau}^{p,q}$ and

$$\begin{aligned} & \int_{I_m} ((U', \varphi) + A_{h,m}(U, U, \varphi) + b_{h,m}(U, \varphi)) \, dt + (\{U\}_{m-1}, \varphi_{m-1}^+) \\ &= \int_{I_m} \ell_{h,m}(\varphi) \, dt, \quad \forall \varphi \in S_{h,\tau}^{p,q}, \quad m = 1, \dots, M, \quad U_0^- := \Pi_1 u^0. \end{aligned} \quad (17)$$

The exact regular solution u satisfies the identity

$$\begin{aligned} & \int_{I_m} ((u', \varphi) + A_{h,m}(u, u, \varphi) + b_{h,m}(u, \varphi)) \, dt + (\{u\}_{m-1}, \varphi_{m-1}^+) \\ &= \int_{I_m} \ell_{h,m}(\varphi) \, dt \quad \forall \varphi \in S_{h,\tau}^{p,q}, \quad \text{with } u(0-) = u(0). \end{aligned} \quad (18)$$

It is also possible to consider $q = 0$. In this case, scheme (17) represents a version of the backward Euler method. Therefore, we shall be concerned only with $q \geq 1$.

3 Error Analysis

In the derivation of the error we shall use the $S_{h,\tau}^{p,q}$ -interpolation π of functions $v \in H^1(0, T; L^2(\Omega))$ defined by

$$\begin{aligned} & \text{a) } \pi v \in S_{h,\tau}^{p,q}, \quad \text{b) } (\pi v)(t_m-) = \Pi_m v(t_m-), \\ & \text{c) } \int_{I_m} (\pi v - v, \varphi^*) \, dt = 0 \quad \forall \varphi^* \in S_{h,\tau}^{p,q-1}, \quad \forall m = 1, \dots, M. \end{aligned} \quad (19)$$

It is possible to prove that πu is uniquely determined and $\pi v|_{I_m} = \pi(\Pi_m v)|_{I_m}$.

Our main goal will be the analysis of the estimation of the error $e = U - u$, which can be expressed in the form $e = \xi + \eta$, where $\xi = U - \pi u \in S_{h,\tau}^{p,q}$ and $\eta = \pi u - u$. Then, in virtue of (17) and (18), for each $\varphi \in S_{h,\tau}^{p,q}$ we have

$$\begin{aligned} & \int_{I_m} ((\xi', \varphi) + A_{h,m}(U, U, \varphi) - A_{h,m}(u, u, \varphi)) dt + (\{\xi_{m-1}\}, \varphi_{m-1}^+) \quad (20) \\ &= \int_{I_m} (b_{h,m}(u, \varphi) - b_{h,m}(U, \varphi)) dt - \int_{I_m} (\eta', \varphi) dt - (\{\eta\}_{m-1}, \varphi_{m-1}^+). \end{aligned}$$

3.1 Derivation of an Abstract Error Estimate

In our further considerations, by C we shall denote a positive generic constant, independent of h, τ, m, M, K, u, U , which can attain different values in different places. In the sequel, we shall consider a system of triangulations $\mathcal{T}_{h,m}$, $m = 1, \dots, M$, $h \in (0, h_0)$, which is shape regular and locally quasuniform: There exist positive constants C_R and C_Q , independent of K, Γ, m, M and h , such that for all $m = 1, \dots, M$ and $h \in (0, h_0)$

$$\frac{h_K}{\rho_K} \leq C_R, \quad \forall K \in \mathcal{T}_{h,m}, \quad (21)$$

$$h_{K_\Gamma}^{(L)} \leq C_Q h_{K_\Gamma}^{(R)}, \quad h_{K_\Gamma}^{(R)} \leq C_Q h_{K_\Gamma}^{(L)} \quad \forall \Gamma \in \mathcal{F}_{h,m}^I. \quad (22)$$

Important tools in the analysis of the DGFEM are the multiplicative trace inequality and the inverse inequality: There exist constants $C_M, C_I > 0$ independent of $h \in (0, h_0), m, M, K \in \mathcal{T}_{h,m}$ and v such that

$$\|v\|_{L^2(\partial K)}^2 \leq C_M \left(\|v\|_{L^2(K)} \|v\|_{H^1(K)} + h_K^{-1} \|v\|_{L^2(K)}^2 \right), \quad v \in H^1(K), \quad (23)$$

and

$$\|v\|_{H^1(K)} \leq C_I h_K^{-1} \|v\|_{L^2(K)}, \quad v \in P^p(K). \quad (24)$$

The analysis of the form $b_{h,m}$ implies that for each $k > 0$ there exists a constant $C = C(k)$ such that

$$\begin{aligned} & |b_{h,m}(U, \varphi) - b_{h,m}(u, \varphi)| \quad (25) \\ & \leq \frac{\beta_0}{k} \|\varphi\|_{DG,m}^2 + C(\|\xi\|^2 + \|\eta\|_{L^2(\Omega)}^2 + \sum_{K \in \mathcal{T}_{h,m}} h_K^2 |\eta|_{H^1(K)}^2). \end{aligned}$$

As for the coercivity, we can prove the following result: Let

$$C_W > 0, \quad \text{for } \theta = -1 \text{ (NIPG)}, \quad (26)$$

$$C_W \geq \left(\frac{4\beta_1}{\beta_0} \right)^2 C_{MI} \quad \text{for } \theta = 1 \text{ (SIPG)}, \quad (27)$$

$$C_W \geq 2 \left(\frac{2\beta_1}{\beta_0} \right)^2 C_{MI} \quad \text{for } \theta = 0 \text{ (IIPG)}, \quad (28)$$

where $C_{MI} = C_M(C_I + 1)(C_Q + 1)$. Then

$$a_{h,m}(U, \xi, \xi) + \beta_0 J_{h,m}(\xi, \xi) \geq \frac{\beta_0}{2} \|\xi\|_{DG,m}^2. \quad (29)$$

Let us substitute $\varphi := \xi$ in (20). Then a detailed technical analysis yields the estimate

$$\begin{aligned} & \|\xi_m^-\|^2 - \|\xi_{m-1}^-\|^2 + \frac{\beta_0}{2} \int_{I_m} \|\xi\|_{DG,m}^2 dt \\ & \leq C \int_{I_m} \|\xi\|^2 dt + 4 \|\eta_{m-1}^-\|^2 + C \int_{I_m} R_m(\eta) dt, \end{aligned} \quad (30)$$

where

$$R_m(\eta) = \|\eta\|_{DG,m}^2 + \|\eta\|^2 + \sum_{K \in \mathcal{T}_{h,m}} (h_K^2 |\eta|_{H^1(K)}^2 + h_K^2 |\eta|_{H^2(K)}^2). \quad (31)$$

An important task is the estimation of the term $\int_{I_m} \|\xi\|^2 dt$. The case, when $\beta(u) = \text{const} > 0$, was analyzed in [6] and [7] using the approach from [1] based on the application of the so-called Gauss-Radau quadrature and interpolation. However, in the case of nonlinear diffusion, this technique is not applicable.

Lemma 1. *There exist constants $C, C^* > 0$ such that*

$$\int_{I_m} \|\xi\|^2 dt \leq C \tau_m \left(\|\xi_{m-1}^-\|^2 + \|\eta_{m-1}^-\|^2 + \int_{I_m} R_m(\eta) dt \right), \quad (32)$$

provided

$$0 < \tau_m \leq C^* \beta_0. \quad (33)$$

Proof. The proof is rather technical. Therefore, we can mention only the most important steps. Let us set

$$t_{m-l/q} = t_{m-1} + \frac{l}{q}(t_m - t_{m-1}) \quad \text{for } l = 0, \dots, q.$$

Using scaling arguments and the equivalence of norms in the space $P^q(0, 1)$, we get the inequalities

$$\sum_{l=0}^q \|\xi(t_{m-l/q})\|^2 \geq \frac{L_q}{\tau_m} \int_{I_m} \|\xi\|^2 dt. \quad (34)$$

and

$$\|\xi_{m-1}^+\|^2 \leq \frac{M_q}{\tau_m} \int_{I_m} \|\xi\|^2 dt \quad (35)$$

with constants L_q, M_q depending on q only.

Let us substitute $\varphi := \xi$ in (20). Then a detailed analysis yields the estimate

$$\begin{aligned} & \|\xi_m^-\|^2 + \|\xi_{m-1}^+\|^2 + \frac{\beta_0}{2} \int_{I_m} \|\xi\|_{DG,m}^2 dt \\ & \leq C \left(\int_{I_m} \|\xi\|^2 dt + \int_{I_m} R_m(\eta) dt \right) + 2 \frac{\|\eta_{m-1}^-\|^2}{\delta_1} + 2 \frac{\|\xi_{m-1}^-\|^2}{\delta_1} + 4\delta_1 \|\xi_{m-1}^+\|^2, \end{aligned} \quad (36)$$

valid for any $\delta_1 > 0$.

In the case $q = 1$, using (34) – (36) and choosing δ_1 in a suitable way, we conclude that Lemma 1 holds.

Further, let $q \geq 2$. For each $l = 1, \dots, q-1$ we set $\tilde{\xi}_l = \zeta_{t_{m-l/q}}$, where $\zeta_{t_{m-l/q}}$ is the discrete characteristic function to the function ξ at the point $t_{m-l/q}$. This means that $\tilde{\xi}_l \in S_{h,\tau}^{p,q}$,

$$\int_{I_m} (\tilde{\xi}_l, \varphi) dt = \int_{t_{m-1}}^{t_{m-l/q}} (\xi, \varphi) dt, \quad \forall \varphi \in S_{h,m}^{p,q-1}, \quad \tilde{\xi}_l(t_{m-1}^+) = \xi(t_{m-1}^+). \quad (37)$$

It is possible to show that

$$\int_{I_m} \|\tilde{\xi}_l\|_{DG,m}^2 dt \leq C \int_{I_m} \|\xi\|_{DG,m}^2 dt. \quad (38)$$

Using in (37) $\varphi := \xi'$, we find that

$$\int_{I_m} (\xi', \tilde{\xi}_l) dt + \left(\xi_{m-1}^+, (\tilde{\xi}_l)_{m-1}^+ \right) = \frac{1}{2} (\|\xi(t_{m-l/q})\|^2 + \|\xi_{m-1}^+\|^2). \quad (39)$$

Using (20) with $\varphi = \tilde{\xi}_l$, (34), (35), (38) and (39), after a detailed computation we find that for any $\delta_2 > 0$ we have

$$\begin{aligned} & \|\xi_{m-l/q}\|^2 + \|\xi_{m-1}^+\|^2 \\ & \leq C \int_{I_m} (\|\xi\|_{DG,m}^2 + \|\xi\|^2 + R_m(\eta)) dt + 2 \frac{\|\xi_{m-1}^-\|^2}{\delta_2} + 2 \frac{\|\eta_{m-1}^-\|^2}{\delta_2} + 4\delta_2 \|\xi_{m-1}^+\|^2. \end{aligned} \quad (40)$$

If we sum (40) over all $l = 1, \dots, q-1$, use (30), (34), (35) and choose δ_2 in a suitable way, we prove the existence of a constant $C^* > 0$ such that (32) holds, if (33) is satisfied.

On the basis of (30) and (32), discrete Gronwall's lemma and the relations $\xi_0^- = 0$, $e = \xi + \eta$ we obtain the abstract error estimate:

Theorem 1. *Let (33) hold. Then there exists a constants $C > 0$ such that the error $e = U - u$ satisfies the estimate*

$$\begin{aligned} & \|e_m^-\|^2 + \frac{\beta_0}{2} \sum_{j=1}^m \int_{I_j} \|e\|_{DG,j}^2 dt \\ & \leq C \left(\sum_{j=1}^m \|\eta_j^-\|^2 + \sum_{j=1}^m \int_{I_j} R_j(\eta) dt \right) + 2 \|\eta_m^-\|^2 + \beta_0 \sum_{j=1}^m \int_{I_j} \|\eta\|_{DG,j}^2 dt, \\ & m = 1, \dots, M, \quad h \in (0, h_0). \end{aligned} \quad (41)$$

3.2 Error Estimation in Terms of h and τ

The derivation of error estimates in dependence on h and τ is obtained from the abstract error estimate and estimation of terms containing η , under the assumptions (7) and

$$u \in H^{q+1}(0, T; H^1(\Omega)) \cap C([0, T]; H^{p+1}(\Omega)), \quad (42)$$

and the assumption that the meshes satisfy conditions (21), (22), (33) and

$$\tau_m \geq Ch_m^2, \quad m = 1, \dots, M. \quad (43)$$

Moreover, we assume that the Dirichlet datum u_D satisfies the condition

$$u_D(x, t) = \sum_{j=0}^q \psi_j(x) t^j, \quad (44)$$

where $\psi_j \in H^{p+1/2}(\partial\Omega)$ for $j = 0, \dots, q$. If all meshes $\mathcal{T}_{h,m}$ are identical, then condition (43) can be omitted.

Then, using a similar process as in [6] and [7], we obtain the *main result*:

Theorem 2. *Let u be the exact solution of problem (1)–(3) satisfying the regularity conditions (7) and (42). Let U be the approximate solution to problem (1)–(3) obtained by scheme (17) in the case that the Dirichlet datum u_D is defined by (44). Let conditions (21), (22), (33) and (43) be satisfied. Then there exists a constant $C > 0$ independent of $h, \tau, m, \varepsilon, u, U$ such that*

$$\begin{aligned} & \|e_m^-\|^2 + \frac{\varepsilon}{2} \sum_{j=1}^m \int_{I_j} \|e\|_{DG,j}^2 dt \\ & \leq C \left(h^{2p} |u|_{C([0,T];H^{p+1}(\Omega))}^2 + \tau^{2q+2} |u|_{H^{q+1}(0,T;H^1(\Omega))}^2 \right), \\ & \quad m = 1, \dots, M, \quad h \in (0, h_0). \end{aligned} \quad (45)$$

The detailed analysis will be a subject of a paper [3] in preparation.

4 DGFEM for the Solution of Compressible Flow in Time-Dependent Domains

4.1 Continuous Problem in the ALE Form

We shall be concerned with the numerical solution of compressible flow in a bounded domain $\Omega_t \subset \mathbb{R}^2$ depending on time $t \in [0, T]$. The time dependence of the domain is taken into account with the aid of a regular one-to-one ALE mapping $\mathcal{A}_t : \overline{\Omega}_0 \longrightarrow \overline{\Omega}_t$. We define the ALE velocity $\tilde{z}(X, t) = \partial \mathcal{A}_t(X) / \partial t$, $z(x, t) = \tilde{z}(\mathcal{A}_t^{-1}(x), t)$, $t \in [0, T]$, $X \in \overline{\Omega}_0$, $x \in \overline{\Omega}_t$, and the ALE derivative of a function $f = f(x, t)$ defined for $x \in \Omega_t$ and $t \in (0, T)$: $D^A f(x, t) / Dt = \partial \tilde{f}(X, t) / \partial t$, where $\tilde{f}(X, t) = f(\mathcal{A}_t(X), t)$, $X = \mathcal{A}_t^{-1}(x) \in \Omega_0$.

The system describing compressible flow consisting of the continuity equation, the Navier-Stokes equations, the energy equation and thermodynamical relations can be written in the ALE form

$$\frac{D^A \mathbf{w}}{Dt} + \sum_{s=1}^2 \frac{\partial \mathbf{g}_s(\mathbf{w})}{\partial x_s} + \mathbf{w} \operatorname{div} \mathbf{z} = \sum_{s=1}^2 \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s}, \quad (46)$$

where

$$\begin{aligned} \mathbf{w} &= (w_1, \dots, w_4)^T = (\rho, \rho v_1, \rho v_2, E)^T \in \mathbb{R}^4, \quad \mathbf{g}_i(\mathbf{w}) = \mathbf{f}_i(\mathbf{w}) - z_i \mathbf{w}, \quad (47) \\ \mathbf{f}_i(\mathbf{w}) &= (f_{i1}, \dots, f_{i4})^T = (\rho v_i, \rho v_1 v_i + \delta_{1i} p, \rho v_2 v_i + \delta_{2i} p, (E + p) v_i)^T, \\ \mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w}) &= (R_{i1}, \dots, R_{i4})^T = (0, \tau_{i1}^V, \tau_{i2}^V, \tau_{i1}^V v_1 + \tau_{i2}^V v_2 + k \partial \theta / \partial x_i)^T, \\ \tau_{ij}^V &= \lambda \operatorname{div} \mathbf{v} \delta_{ij} + 2\mu d_{ij}(\mathbf{v}), \quad d_{ij}(\mathbf{v}) = (\partial v_i / \partial x_j + \partial v_j / \partial x_i) / 2, \\ p &= (\gamma - 1)(E - \rho |\mathbf{v}|^2 / 2), \quad \theta = (E / \rho - |\mathbf{v}|^2 / 2) g / c_v. \end{aligned} \quad (48)$$

We use the following notation: ρ - density, p - pressure, E - total energy, $\mathbf{v} = (v_1, v_2)$ - velocity, θ - absolute temperature, $\gamma > 1$ - Poisson adiabatic constant, $c_v > 0$ - specific heat at constant volume, $\mu > 0, \lambda = -2\mu/3$ - viscosity coefficients, $k > 0$ - heat conduction.

The above system is equipped with initial condition

$$\mathbf{w}(x, 0) = \mathbf{w}^0(x), \quad x \in \Omega_0. \quad (49)$$

As for boundary conditions, we assume that the boundary of Ω_t consists of three different parts: $\partial \Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$, where Γ_I is the inlet, Γ_O is the outlet and Γ_{W_t} denotes impermeable walls that may move in dependence on time. Then we prescribe the following boundary conditions:

$$\text{a) } \rho|_{\Gamma_I} = \rho_D, \quad \text{b) } \mathbf{v}|_{\Gamma_I} = \mathbf{v}_D = (v_{D1}, v_{D2})^T, \quad (50)$$

$$\text{c) } \sum_{i,j=1}^2 \tau_{ij}^V n_i v_j + k \frac{\partial \theta}{\partial n} = 0 \quad \text{on } \Gamma_I,$$

$$\text{a) } \mathbf{v}|_{\Gamma_{W_t}} = \mathbf{z}_D = (z_{D1}, z_{D2}), \quad \text{b) } \frac{\partial \theta}{\partial n}|_{\Gamma_{W_t}} = 0, \quad (51)$$

$$\text{a) } \sum_{i=1}^2 \tau_{ij}^V n_i = 0, \quad j = 1, 2, \quad \text{b) } \frac{\partial \theta}{\partial n} = 0 \quad \text{on } \Gamma_O. \quad (52)$$

By $\mathbf{z}_D$ we denote the velocity of a moving wall.

4.2 Discretization

Let us construct a partition $0 = t_0 < t_1 < t_2 \dots$ of the time interval $[0, T]$. At each time instant t_m , the domain Ω_{t_m} is approximated by a polygonal domain $\Omega_{h,m}$, in which a triangulation $\mathcal{T}_{h,m}$ is constructed. The discrete problem is formulated in a similar way as in Section 2. The approximate solution will be

denoted by $\mathbf{W}$. We assume that $\mathbf{W}|_{I_m} \in \mathcal{S}_{h,\tau,m}^{p,q} = \{\boldsymbol{\varphi} \in L^2(\Omega_{h,m} \times I_m); \boldsymbol{\varphi} = \sum_{i=0}^q t^i \boldsymbol{\varphi}_i \text{ with } \boldsymbol{\varphi}_i \in [S_{h,m}^p]^4, t \in I_m\}$. The symbol $\mathcal{F}_{h,m}^D$ will denote the system of $\Gamma \in \mathcal{F}_{h,m}^B$, on which a Dirichlet condition is prescribed. We introduce the forms

$$a_{h,m}(\mathbf{w}, \boldsymbol{\varphi}_h) = \sum_{K \in \mathcal{T}_{h,m}} \int_K \sum_{s=1}^2 \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} dx \quad (53)$$

$$\begin{aligned} & - \sum_{\Gamma \in \mathcal{F}_{h,m}^I} \int_{\Gamma} \sum_{s=1}^2 \langle \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) \rangle (\mathbf{n}_{\Gamma})_s \cdot [\boldsymbol{\varphi}_h] dS \\ & - \sum_{\Gamma \in \mathcal{F}_{h,m}^D} \int_{\Gamma} \sum_{s=1}^2 \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) (\mathbf{n}_{\Gamma})_s \cdot \boldsymbol{\varphi}_h dS, \end{aligned} \quad (54)$$

$$\begin{aligned} b_{h,m}(\mathbf{w}, \boldsymbol{\varphi}_h) &= - \sum_{K \in \mathcal{T}_{h,m}} \int_K \sum_{s=1}^2 \mathbf{g}_s(\mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} dx \\ &+ \sum_{\Gamma \in \mathcal{F}_{h,m}^I} \int_{\Gamma} \mathbf{H}_g(\mathbf{w}_{\Gamma}^{(L)}, \mathbf{w}_{\Gamma}^{(R)}, \mathbf{n}_{\Gamma}) \cdot [\boldsymbol{\varphi}_h] dS \\ &+ \sum_{\Gamma \in \mathcal{F}_{h,m}^B} \int_{\Gamma} \mathbf{H}_g(\mathbf{w}_{\Gamma}^{(L)}, \mathbf{w}_{\Gamma}^{(R)}, \mathbf{n}_{\Gamma}) \cdot \boldsymbol{\varphi}_h dS, \end{aligned} \quad (55)$$

$$\begin{aligned} J_{h,m}(\mathbf{w}, \boldsymbol{\varphi}_h) &= \sum_{\Gamma \in \mathcal{F}_{h,m}^I} \int_{\Gamma} h(\Gamma)^{-1} [\mathbf{w}] \cdot [\boldsymbol{\varphi}_h] dS \\ &+ \sum_{\Gamma \in \mathcal{F}_{h,m}^D} \int_{\Gamma} h(\Gamma)^{-1} \mathbf{w} \cdot \boldsymbol{\varphi}_h dS, \end{aligned} \quad (56)$$

$$\ell_{h,m}(\mathbf{w}, \boldsymbol{\varphi}_h) = \sum_{\Gamma \in \mathcal{F}_{h,m}^D} \int_{\Gamma} \sum_{s=1}^2 h(\Gamma)^{-1} \mathbf{w}_B \cdot \boldsymbol{\varphi}_h dS, \quad (57)$$

$$d_{h,m}(\mathbf{w}, \boldsymbol{\varphi}_h) = \sum_{K \in \mathcal{T}_{h,m}} \int_K (\mathbf{w} \cdot \boldsymbol{\varphi}_h) \operatorname{div} \mathbf{z} dx. \quad (58)$$

Here $\mathbf{H}_g$ is a conservative numerical flux consistent with the fluxes $\mathbf{g}_s$. We use the incomplete IIPG version (i.e. $\theta = 0$). The boundary state $\mathbf{w}_B$ is defined on the basis the Dirichlet boundary conditions and extrapolation. For $\Gamma \in \mathcal{F}_{h,m}^B$ the boundary state $\mathbf{w}_{\Gamma}^{(R)}$ appearing in the form $b_{h,m}$ is defined with the aid of the solution of the 1D linearized initial-boundary Riemann problem as in [4].

Further, we set $\widehat{\mathbf{W}}_{m-1}^- = \mathbf{W}_{m-1}^- \circ \mathcal{A}_{t_{m-1}}^{-1} \circ \mathcal{A}_{t_m}$. Now we can define the approximate solution as a function $\mathbf{W}$ satisfying the conditions $\mathbf{W}|_{I_m} \in \mathcal{S}_{h,\tau,m}^{p,q}$ and

$$\int_{I_m} ((\mathbf{W}', \boldsymbol{\varphi}) + a_{h,m}(\mathbf{W}, \boldsymbol{\varphi}) + b_{h,m}(\mathbf{W}, \boldsymbol{\varphi}) + J_{h,m}(\mathbf{W}, \boldsymbol{\varphi}) \quad (59)$$

$$+d_{h,m}(\mathbf{W}, \varphi)) dt + \left(\mathbf{W}_{m-1}^+ - \widehat{\mathbf{W}}_{m-1}^-, \varphi_{m-1}^+ \right) = \int_{I_m} \ell_{h,m}(\varphi) dt,$$

$$\forall \varphi \in \mathcal{S}_{h,\tau,m}^{p,q}, \quad m = 1, \dots, M, \quad \mathbf{W}_0^- := \Pi_1 u^0.$$

This nonlinear problem is solved with respect to $\mathbf{W}|_{I_m}$ by a suitable iterative process.

4.3 Flow Induced Airfoil Vibrations

We consider an elastically supported airfoil with two degrees of freedom - the vertical displacement H (positively oriented downwards) and the angle α of rotation around an elastic axis EO (positively oriented clockwise). The motion of the airfoil is described by the system of nonlinear ordinary differential equations for unknowns H, α :

$$m\ddot{H} + k_{HH}H + S_\alpha \ddot{\alpha} \cos \alpha - S_\alpha \dot{\alpha}^2 \sin \alpha + d_{HH}\dot{H} = -L(t), \quad (60)$$

$$S_\alpha \ddot{H} \cos \alpha + I_\alpha \ddot{\alpha} + k_{\alpha\alpha}\alpha + d_{\alpha\alpha}\dot{\alpha} = M(t).$$

We use the following notation: m - mass of the airfoil, S_α - static moment around the elastic axis $EO = (x_{EO1}, x_{EO2})$, I_α - inertia moment around the elastic axis EO , k_{HH} - bending stiffness, $k_{\alpha\alpha}$ - torsional stiffness, d_{HH} - structural damping in bending, $d_{\alpha\alpha}$ - structural damping in torsion, c - length of the chord of the airfoil, l - airfoil depth. The aerodynamic lift force L and aerodynamic torsional moment M are defined by

$$L = -l \int_{\Gamma_{Wt}} \sum_{j=1}^2 \tau_{2j} n_j dS, \quad M = l \int_{\Gamma_{Wt}} \sum_{i,j=1}^2 \tau_{ij} n_j r_i^{\text{ort}} dS, \quad (61)$$

$$\tau_{ij} = -p\delta_{ij} + \tau_{ij}^V, \quad r_1^{\text{ort}} = -(x_2 - x_{EO2}), \quad r_2^{\text{ort}} = x_1 - x_{EO1}.$$

System (60) is equipped with the initial conditions prescribing the values $H(0)$, $\alpha(0)$, $\dot{H}(0)$, $\dot{\alpha}(0)$. It is transformed to a first-order ODE system and approximated by the fourth-order Runge-Kutta method coupled with scheme (59).

Figure 1 shows the displacement H and the rotation angle α in dependence on time for the far-field velocity 20, 30 and 40 m/s and the following data: $m = 0.086622$ kg, $S_a = -0.000779673$ kg m, $I_a = 0.000487291$ kg m², $k_{HH} = 105.109$ N/m, $k_{\alpha\alpha} = 3.696682$ Nm/rad, $l = 0.05$ m, $c = 0.3$ m, $\mu = 1.8375 \cdot 10^{-5}$ kg m⁻¹ s⁻¹, far-field density $\rho = 1.225$ kg m⁻³, $H(0) = -0.02$ m, $\alpha(0) = 6$ degrees, $\dot{H}(0) = 0$, $\dot{\alpha} = 0$. The position of the elastic axis is on the chord of the airfoil at the 40% distance from the leading edge. The far-field Mach number is 0.014 for the velocity 20 m/s. The structural damping is neglected. The flow is purely subsonic in this case and, therefore, it is not necessary to introduce an artificial viscosity in scheme (59), as was carried out, e.g. in [8]. In (59), the approximation polynomial degrees $q = 0, p = 2$ were used. We see that for the velocities 20 and 30 m/s the vibrations are damped, but for the velocity 40 m/s we get the flutter instability when the vibration amplitudes are increasing in

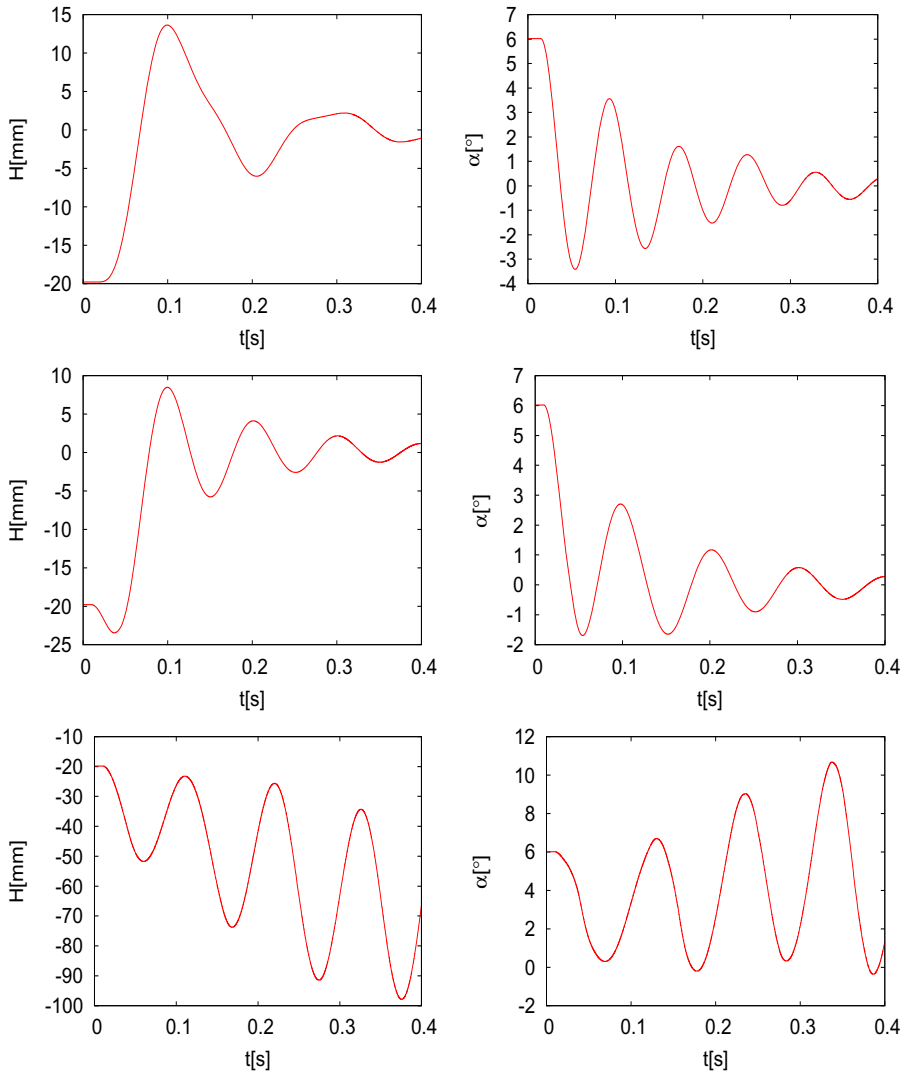


Fig. 1. Displacement H (left) and rotation angle α (right) of the airfoil in dependence on time for far-field velocity 20, 30 and 40 m/s

time. The monotonous increase and decrease of the average values of H and α , respectively, shows that the flutter is combined with a divergence instability in the presented example.

Acknowledgements. This work is supported by the research project MSM 0021620839 (M. Feistauer) and by the Nečas Center for Mathematical Modelling, project LC06052 (J. Česenek), both financed by the Ministry of Education of the Czech Republic. The research of J. Česenek was also partly supported by the project No. 12810 of the Grant Agency of the Charles University in Prague.

References

1. Akrivis, G., Makridakis, C.: Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: Math. Modelling and Numer. Anal.* 38, 261–289 (2004)
2. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* 39, 1749–1779 (2001)
3. Česenek, J., Feistauer, M.: Theory of the space-time discontinuous Galerkin method for nonstationary parabolic problems with nonlinear convection and diffusion (in preparation)
4. Feistauer, M., Česenek, J., Horáček, J., Kučera, V., Prokopová, J.: DGFEM for the numerical solution of compressible flow in time dependent domains and applications to fluid-structure interaction. In: Pereira, J.C.F., Sequeira, A. (eds.) *Proceedings of the 5th European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, Lisbon, Portugal, June 14–17 (2010) (published electronically), ISBN 978-989-96778-1-4
5. Feistauer, M., Hájek, J., Švadlenka, K.: Space-time discontinuous Galerkin method for solving nonstationary linear convection-diffusion-reaction problems. *Appl. Math.* 52, 197–234 (2007)
6. Feistauer, M., Kučera, V., Najzar, K., Prokopová, J.: Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. Preprint No. MATH-knm-2010/1, Charles University Prague, School of Mathematics (submitted *Numer. Math.*)
7. Feistauer, M., Kučera, V., Najzar, K., Prokopová, J.: Space-time DG method for nonstationary convection-diffusion problems. In: *Numerical Mathematics and Advanced Applications, ENUMATH 2009*. Springer, Heidelberg (2010), doi:10.1007/978-3-642-11795-4_34
8. Feistauer, M., Kučera, V., Prokopová, J.: Discontinuous Galerkin solution of compressible flow in time-dependent domains. *Mathematics and Computers in Simulations* 80, 1612–1623 (2010)
9. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: *Computational Differential Equations*. Cambridge University Press, Cambridge (1996)
10. Thomée, V.: *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin (2006)

Stochastic Algorithms in Linear Algebra - beyond the Markov Chains and von Neumann - Ulam Scheme*

Karl Sabelfeld

Institute Comp. Math. & Math. Geoph., Novosibirsk, Lavrentiev str, 6,
630090 Novosibirsk, Russia
sabelfeld.karl@yahoo.de

Abstract. Sparsified Randomization Monte Carlo (SRMC) algorithms for solving systems of linear algebraic equations introduced in our previous paper [34] are discussed here in a broader context. In particular, I present new randomized solvers for large systems of linear equations, randomized singular value (SVD) decomposition for large matrices and their use for solving inverse problems, and stochastic simulation of random fields. Stochastic projection methods, which I call here "random row action" algorithms, are extended to problems which involve systems of equations and constraints in the form of systems of linear inequalities.

1 Introduction

The use of Monte Carlo methods for solving large systems of linear equations is intimately tied the Neumann-Ulam scheme, e.g., see [15], [16], [20], [37], [31], [32], [5], [6], [7]. It can be interpreted as follows: (1) first, take the representation of the solution in a form of the Neumann series, then, (2) represent the solution (one component of the vector, in the case of a system of algebraic equations $x = Ax + b$) as an expectation over some Markov chain associated in a sense to the matrix A , (3) the expectation is then calculated by taking an ensemble average (numerically, the arithmetic mean) of a random estimator defined on the constructed Markov chains.

The nice feature of this method has always its parsimonious memory usage: the method takes almost no memory, independent of the size of the matrix. However a serious drawback is its weak convergence: the error decreases as $O(N^{-1/2})$ where N is the number of independent samples of the Markov chains. Quasi-Monte Carlo methods may sometimes improve the rate of convergence, however in practice the improvement is often too small.

Nowadays, there has been dramatic progress in solving the storage problem, and it is natural to involve other stochastic ideas beyond the von

* The author thanks the organizers of the conference, and acknowledges the support of the RFBR under Grants N 06-01-00498, 09-01-12028-ofi-m, and a joint BMBF and Bortnik Funds Grant N 7326.

Neumann-Ulam-Markov chain paradigm. As an example, we mention conventional deterministic iteration methods where however the weights are chosen at random (e.g., see [42], [36]). Another important example is the projection method where one takes projections onto randomly sampled subspaces (e.g., see [41], [33]). Sampling from random subspaces is the main idea also in the randomized singular value decomposition technique (e.g., see [18], [10]-[12]). A general idea behind these methods appeals to the fundamental result of Johnson and Lindenstrauss [21] which says that any n point subset of Euclidean space can be embedded in $k = O(\log\{n\}/\varepsilon^2)$ dimensions without distorting the distances between any pair of points by more than a factor of $(1 \pm \varepsilon)$, for any $0 < \varepsilon < 1$. So the result of Johnson and Lindenstrauss asserts that any set of n points in d -dimensional Euclidean space can be embedded into k -dimensional Euclidean space where k is logarithmic in n and independent of d so that all pairwise distances are maintained within an arbitrarily small factor. The linear transformation can be done by a random matrix whose entries are independent standard Gaussian random variables.

This transformation was essentially simplified in [1] by showing that this matrix can be changed with a matrix whose entries r_{ij} are independent discrete random variables with the distribution $P(\pm 1) = 1/6$, $P(0) = 1/3$ which greatly sparsifies the matrix. More precisely, Achlioptas' Theorem is formulated as follows.

Suppose that A is an $n \times d$ matrix of n points in $\mathbf{R}^d$. Fix constants $\varepsilon, \beta > 0$, and choose an integer k such that

$$k \geq \frac{4 + 2\beta}{\varepsilon^2 - \varepsilon^3/3} \log n .$$

Suppose that R is a random $k \times d$ matrix with entries r_{ij} belonging to the distribution

$$r_{ij} = \sqrt{3} \begin{cases} +1 & p=1/6 \\ 0 & p=2/3 \\ -1 & p=1/6 \end{cases} , \quad (1)$$

Define the $n \times k$ matrix $Q = \frac{1}{\sqrt{k}} AR$, which is considered as a projection of A onto a k -dimensional subspace. For any row $\mathbf{u}$ in A , let $f(\mathbf{u})$ be the corresponding row in Q . Then, for any distinct rows $\mathbf{u}, \mathbf{v}$ of A , we have

$$(1 - \varepsilon) \|\mathbf{u} - \mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1 + \varepsilon) \|\mathbf{u} - \mathbf{v}\|^2$$

with probability at least $1 - n^{-\beta}$.

In [2], the authors suggested a low-distortion embedding of L_2^d into $L_p^{O(\log n)}$ ($p = 1, 2$), called the Fast-Johnson-Lindenstrauss- Transform (FJLT). The FJLT is faster than standard random projections and just as easy to implement. It is based upon the preconditioning of a sparse projection matrix with a randomized Fourier transform.

In all this methods we deal with conventional numerical methods, but introduce some randomness to improve the convergence and more, to turn to

very high dimensions which can not be treated pure deterministically. So for instance, it is well known that the computational cost of a full SVD for large matrices is rapidly increasing with the matrix dimension. The randomized SVD solves this problem by a random sampling of small size submatrices for which the SVD is computed. In the case of projection methods, one projects the points only to a random set of subspaces. This type of methods treats the dimension problem in a non-trivial manner. But the main advantage of these methods is in their convergence rate: it is dramatically increased compared to the conventional Monte Carlo methods, and is actually comparable with the best deterministic methods.

The computational cost of most simulation algorithms in dimension m is increasing exponentially in m . Note that even simply accessing a vector in dimension m requires N^m operations, where N is the number of entries in each direction. This complexity growth is often mentioned as *Curse of Dimensionality* [4]. Given an equation in m dimensions, one can try to approximate its solution $u(x_1, \dots, x_m)$ by a separable function, say, as $u(x_1, \dots, x_m) \approx f_1(x_1) \dots f_m(x_m)$, hence, radically reducing the complexity to a linear function of m . More generally, a *separation representation* is defined as [4]

$$u(x_1, \dots, x_m) = \sum_{i=1}^s \lambda_i f_1^{(i)}(x_1) \dots f_m^{(i)}(x_m) + O(\varepsilon) .$$

Setting an accuracy goal ε first, and then adapting $\{\lambda_i\}$, $\{f_i(x_i)\}$ and s to achieve this goal with minimal separation rank s is the idea behind many algorithms based on the separation representation approach.

In Monte Carlo methods, one often has to deal with very large dimensions, in problems like the integration, solution of integral equations, PDEs, simulation of random fields, etc. It is customary to think that the Monte Carlo methods are able to resolve problems for very high dimensions, however it is true only under the following conditions: (1) the variance of the MC estimator is small, (2) the desired accuracy is not high, (3) the complexity of construction of the random estimator is a slow function of the dimension m . Condition (3) can be often satisfied, however the conditions (1) and (2) are the main concern, because the convergence rate of MC methods is slow, scaling as $\sigma/\sqrt{M}$ where σ is the standard deviation, and M is the sample size. Therefore, any approach, method or algorithm capable to influence one of the above three conditions is of great interest in Monte Carlo methods. In particular, one often says, in a very general sense, that a *variance reduction* is developed when certain deterministic transformations lead to transformed random estimator with smaller variance. The dimension is of less concern, though the dimension reduction is desirable in relation, again, with the variance reduction. For instance, the variance is reduced if an exact (or an efficient deterministic) integration over a part of variables is possible.

In linear algebra, a fundamental approach to separation representations for matrices is based on SVD [19], see also an excellent tutorial presentation [38]. The literature on the numerical construction of SVD is vast, we mention only some

of them, e.g., [19], [24], [26], [29], [40], [44]. Recently, different matrix operations like matrix multiplication and SVD for large matrices based on randomization idea has been suggested in different papers, for different application fields, e.g., see [18], [8], [9], [10], [11], [12], [34], [13], [23], [4], [44], [25].

Where can these computational techniques be employed? Essentially in all fields where computation is extensively used, especially when dealing with very high dimensions, such as with high-dimensional PDEs, integral equations of the 3D potential theory, inverse problems of tomography and crystallography, solving the Schrödinger equation, turbulence simulations. These techniques prove useful not only in the computational mathematics, but also problems from information retrieval and Web analysis, such as Google PageRank problem and latent semantic indexing, have strongly motivated the research in the field of design and analysis of linear algebra algorithms involving massive data sets. The list of applications can be easily extended by Data clustering, information retrieval, property testing of graphs, image processing, among others.

2 Sparsified Randomization Algorithms for Linear Systems

Let us consider a system of linear algebraic equations with a $n \times n$ matrix A ,

$$\mathbf{x} = A\mathbf{x} + \mathbf{b}, \quad (2)$$

$\mathbf{x} = (x_1, \dots, x_n)^T$, $\mathbf{b} = (b_1, \dots, b_n)^T \in R^n$, and $A = \{A_{ij}\}_{i,j=1}^n$, where ^{T} stands for the transpose operation, and n is supposed to be large enough.

For simplicity, we assume that the spectral radius of the matrix A is less than unity, so that the solution of (2) can be calculated by the simple iteration method

$$\mathbf{x}^{(m+1)} = A\mathbf{x}^{(m)} + \mathbf{b}; \quad \mathbf{x}^{(0)} = \mathbf{b}; \quad m = 0, 1, 2, \dots \quad (3)$$

Generalizations to other iteration methods are presented in our paper [34].

Sampling of Columns without Replacement. Let G be an unbiased estimator for the matrix A which is defined as a random matrix such that $EG = A$, and let $G^{(0)}, G^{(1)}, \dots, G^{(M-1)}$ be a sequence of independent samples chosen from the random estimator G . The iterative procedure is defined by

$$\boldsymbol{\xi}^{(m+1)} = G^{(m)}\boldsymbol{\xi}^{(m)} + \mathbf{b}, \quad m = 0, 1, \dots, M-1 \quad (4)$$

where $\boldsymbol{\xi}^{(0)} = \mathbf{b}$. Since $G^{(m)}$, $m = 0, 1, \dots$ are all independent of each other, we get from (4) that $\mathbf{E}\boldsymbol{\xi}^{(M)} = \mathbf{x}^{(M)}$.

Let us consider the particular case when G is chosen as a sparse matrix. We will construct the matrix G column-wise: fix an arbitrary integer l which is much less than n , and choose a random set $\mathbf{J}$ of l integers uniformly from 1 to n without replacement, that is, we choose j_1 as an integer uniformly among

$1, 2, \dots, n$, then, j_2 uniformly among the rest of $n - 1$ integers, etc., the last being j_l , and define the entries of G by

$$G_{ik} = \begin{cases} \frac{n}{l} A_{ik} & \text{for } k \in \mathbf{J} \\ 0 & \text{else} \end{cases}$$

for $i = 1, 2, \dots, n$.

Thus, the random matrix G has exactly l nonzero columns of the matrix A , and obviously that for any i, k we have

$$\mathbf{E}G_{ik} = G_{ik} \mathbf{P}\{k \in \mathbf{J}\} = A_{ik}.$$

Note that for calculation of the components of the vector $\boldsymbol{\xi}^{(m+1)}$ we need only l components of the vector $\boldsymbol{\xi}^{(m)}$ and in order to calculate them we need only l components of $\boldsymbol{\xi}^{(m-1)}$, and so on. Consequently, we need l^2 operations in every step. For approximation of $\mathbf{x}^{(M)}$ we need NMl^2 operations, where N is the necessary statistics and M is the length of the cut-off of the Neumann series.

Non-uniform Sampling of Columns with Replacement. Let us present a different version of the sparsification algorithm, where the random choice of columns is not uniform, but it is carried out as a sampling with replacement. In addition, for generality, we describe the evaluation of AB where B is a vector or a matrix.

Starting with the remark that a product of two matrices, A and B , can be represented as follows, $AB = \sum_{\tau=1}^n A_{(\tau)} B^{(\tau)}$ where we use the notation $A_{(\tau)}$ for the τ -th column of A , and $B^{(\tau)}$ for the τ -th row of B we come to the randomized calculation of AB .

Let us choose a probability distribution $p_1, p_2, \dots, p_n$ for sampling from the indices $1, 2, \dots, n$. The randomized evaluation of the product AB is formulated as follows:

1. For $\tau = 1$ to l we sample independently a random number i_τ in $(1, \dots, n)$ according to the probability distribution $\text{Prob}(i_\tau = k) = p_k$, $k = 1, \dots, n$
 - a column of S is chosen as $A_{(i_\tau)}/\sqrt{l p_{i_\tau}}$, and the relevant row in the matrix R is taken as $B^{(i_\tau)}/\sqrt{l p_{i_\tau}}$.
2. The unbiased estimator for AB is the matrix SR .

The estimator SR is obviously unbiased: $E[(SR)_{ij}] = (AB)_{ij}$, $i, j = 1, \dots, n$.

A criterion for the best choice of the distribution $\{p_k\}$ can be of course different. It is convenient to use the mean error in the Frobenius norm, so we have to minimize the quantity $E[\|AB - SR\|_F^2]$.

It can be shown (see [34]) that the choice

$$p_k = \frac{|A_{(k)}| |B^{(k)}|}{\sum_{k=1}^n |A_{(k)}| |B^{(k)}|} \quad (5)$$

minimizes the variance of the error which takes in this case the form:

$$E[||AB - SR||_F^2] = \frac{1}{l} \left(\sum_{k=1}^n |A_{(k)}| |B^{(k)}| \right)^2 - \frac{1}{l} ||AB||_F^2. \quad (6)$$

In conclusion we summarize that in the Sparsified Algorithm we have the following input parameters: n , the size of the matrix A , m , the number of iterations, and l , the size of the sampled submatrices which characterizes how sparse the random matrices in the randomization algorithm are.

3 SVD and Randomized Versions

3.1 SVD Background

Let A be a rectangular $m \times n$ matrix with m rows and n columns, having rank r . From the fundamental theorem of linear algebra we know (e.g., see [38]) that the matrix can be represented as a sum of r matrices of rank 1:

$$A = \sum_{i=1}^r \sigma_i u^{(i)} v^{(i)T} \quad (7)$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ are the singular values, and $u^{(i)} \in \mathbf{R}^m$, $v^{(i)} \in \mathbf{R}^n$, $i = 1, \dots, r$ are its left and right singular column-vectors, respectively. The families $\{u^{(i)}\}$, $\{v^{(i)}\}$ are orthogonal sets of vectors: $u^{(i)T} \cdot u^{(j)} = \delta_{ij}$, and the same for $\{v^{(i)}\}$.

In matrix form, the SVD representation (7) reads:

$$A = U \Sigma V^T \quad (8)$$

where U and V are orthonormal matrices with left and right singular vectors of A , respectively, and Σ is a diagonal matrix: $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$.

Recall that $U^T U = I_{r \times r}$ and $V^T V = I_{n \times n}$. The Frobenius norm $||A||_F$ and the spectral norm $||A||_2$ are defined by

$$||A||_F = \left(\sum_{ij} a_{ij}^2 \right)^{1/2}, \quad ||A||_2 = \max_{|x|_2=1} |Ax|_2 = \sigma_1. \quad (9)$$

The following fundamental result is well known from linear algebra as the *Eckhart-Young theorem* [14]. *If we are interested in the best approximation (in the norms $||\cdot||_F$ and $||\cdot||_2$) of A among all matrices D of rank k , then the solution is $A_k = \sum_{i=1}^k \sigma_i u^{(i)} (v^{(i)})^T$, i.e., for all k rank matrices D , $||A - A_k||_2 \leq ||A - D||_2$, $||A - A_k||_F \leq ||A - D||_F$. The matrix A_k admits the representation:*

$$A_k = U_k \Sigma_k V_k^T = A V_k V_k^T = U_k U_k^T A$$

where U_k, V_k are submatrices of U and V which contain only the top left and right singular vectors, respectively.

A matrix A has a good rank k approximation if $\|A - A_k\|$ is small in Frobenius and 2-norms. To estimate the errors, one may use the well known equalities:

$$\|A - A_k\|_F = \left(\sum_{i=k+1}^r \sigma_i^2(A) \right)^{1/2}, \quad \|A - A_k\|_2 = \sigma_{k+1}(A).$$

3.2 Randomized SVD Algorithm

So let us assume that the matrix A is large enough, and we want to construct a randomized approximation of the first k right singular values and corresponding right singular vectors. The idea behind many versions of randomized algorithms for SVD is to sample randomly s rows of A , then to form an $s \times s$ matrix S and compute its right singular vectors. Let us give the following version presented in [10].

Let us choose a discrete probability distribution $p_1, \dots, p_m$ for sampling from the rows $A_{(1)}, \dots, A_{(m)}$ of A : $\sum_{i=1}^m p_i = 1$.

Randomized SVD Algorithm

0. Fix an integer s such that s is much larger than k , where ε is an error measure, but $s \leq m$.

1. for $j = 1$ to s do:

sample a random index $\{1, \dots, m\}$ of the row of A according to the probability distribution $\{p_j\}_{j=1}^m$, and include $A_{(j)}/\sqrt{sp_j}$ as a row of S ,

2. Compute $S S^T$ and its SVD:

$$S S^T = \sum_{j=1}^s \lambda_j^2 w^{(j)} w^{(j)T}$$

3. Compute $h^{(j)} = S^T w^{(j)} / |S^T w^{(j)}|$ for $j = 1, \dots, k$. Construct H_k as a matrix whose columns are the $h^{(j)}$, and $\lambda_1, \dots, \lambda_k$ are our approximations to the first k singular values of A . Thus we get a rank (at most) k approximation to A is $AH_k H_k^T$.

Note that we could turn to sample columns of A instead of rows, and compute approximations of the left singular vectors, then, H_k were a matrix $RR^T A$ where R is a $m \times k$ matrix containing approximations to the top k left singular vectors.

Let us give the error estimators presented in [12].

Assume that we construct a k rank approximation $AH_k H_k^T$ to our matrix A by the above algorithm where the sampling of s random rows is carried out according to a probability distribution $\{p_i\}_{i=1}^m$ satisfying the condition $p_i \geq \beta |A_{(i)}|^2 / \|A\|_F^2$ for some positive $\beta \leq 1$, and let $\varepsilon > 0$. If $s \geq 4k/\beta\varepsilon^2$ then the following estimation of the mean is true

$$E \left[\|A - AH_k H_k^T\|_F^2 \right] \leq \|A - A_k\|_F^2 + \varepsilon \|A\|_F^2. \quad (10)$$

Error estimation in probability is also possible. Let $\eta = 1 + \sqrt{8 \log(1/\delta)/\beta}$. If $s \geq 4k\eta^2/\beta\varepsilon^2$ then with probability at least $1 - \delta$

$$\|A - AH_k H_k^T\|_F^2 \leq \|A - A_k\|_F^2 + \varepsilon \|A\|_F^2. \quad (11)$$

The same estimations in the spectral norm hold also true, with omitting the factor k in the conditions $s \geq 4k/\beta\varepsilon^2$ and $s \geq 4k\eta^2/\beta\varepsilon^2$.

From the description of the above algorithm it is clear that the steps 1 and 2 are crucial for the efficiency of the method. In the step 1, we could of course use the uniform sampling which means, one call of the RAND generator will be used only, not depending on the dimension n . However this would work well only if the "weights" of the rows, $|A_{(i)}|$ are more or less equal for all $i = 1, \dots, n$. Generally, according to the estimates (10), (11), it is reasonable to sample the rows according to the probability distribution $p_i = \beta |A_{(i)}|^2 / \|A\|_F^2$. In [8], the authors suggest to use the conventional sampling algorithm which needs about $n \log n$ operations. But we can use Walker's algorithm [43] (see the Fortran code in our recent paper [34]) which even in the general case needs only one call to RAND generator, not depending on the dimension of the matrix. Out of the loop, we need only a preparation of two additional arrays of dimension n which are calculated in $O(n)$ operations. This method works of course if we use the sampling of rows with replacement which is always the case since we deal with matrices of large dimension. Thus this sampling algorithm is practically equivalent in efficiency to the uniform sampling of rows !

4 Simulation of Random Fields Based on the Karhunen-Loève Expansion

Let us now consider a real-valued inhomogeneous random field $u(x)$, $x \in G$ defined on a probability space (Ω, A, P) and indexed on a bounded domain G .

Assume (without loss of generality) that the field has a zero mean and a variance $E u^2(x)$ that is bounded for all $x \in G$. The Karhunen-Loève expansion has the form

$$u(x) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k h_k(x), \quad (12)$$

where λ_k and $h_k(x)$ are the eigen-values and eigen-functions of the covariance function $B(x_1, x_2) = \langle u(x_1) u(x_2) \rangle$, and ξ_k is a family of random variables.

Thus λ_k and $h_k(x)$ are the eigen-values and eigen-functions are the solutions of the following eigen-value problem for the correlation operator:

$$\int_G B(x_1, x_2) h_k(x_1) dx_1 = \lambda_k h_k(x_2). \quad (13)$$

The eigen-functions form a complete orthogonal set $\int_G h_i(x) h_j(x) dx = \delta_{ij}$ where δ_{ij} is the Kronecker delta-function. The family $\{\xi_k\}$ is a set of uncorrelated random variables which are obviously related to h_k by

$$\xi_k = \frac{1}{\sqrt{\lambda_k}} \int_G u(x) h_k(x) dx, \quad E \xi_k = 0, \quad E \xi_i \xi_j = \delta_{ij}. \quad (14)$$

It is well known that the Karhunen-Loève expansion presents an optimal (in the mean square sense) convergence for any distribution of $u(x)$. If $u(x)$ is a zero mean Gaussian random field, then $\{\xi_k\}$ is a family of standard Gaussian random variables. Some generalizations to non-gaussian random fields are reported in [27].

4.1 Discrete Approximation of the Karhunen Loève Expansion

Exact solution of the eigen-value problem (13) can be obtained only for some simple cases, but generally, one has to solve it numerically, using quadrature-based methods, e.g., the Nyström method [3].

Assume for simplicity the random process $u(x)$ is defined on a bounded interval $G = [a, b]$, and $x_i, i = 1, \dots, n$ are points of a subdivision of this interval, and we are seeking for a discrete approximation $v \approx u(x)$ where the component v_j of the vector v approximates the value $u(x_j)$, $j = 1, \dots, n$.

Then the covariance $n \times n$ matrix $B_v = \langle v v^T \rangle$ of the vector v should approximate the given correlation function $B(x_i, x_j)$ in the sense that $(B_v)_{ij} \approx B(x_i, x_j)$. This implies that the continuous eigenvalue problem (13) is approximated by the eigenvalue problem for the correlation matrix B_v :

$$B_v g_k = \lambda_k g_k \quad (15)$$

where λ_k are the eigenvalues, and g_k the relevant eigenvectors. Since B_v is symmetric and positive definite, all eigenvalues $\lambda_1, \dots, \lambda_n$ are non-negative, and the spectral representation for the matrix B_v reads

$$B_v = \sum_{k=1}^n \lambda_k g_k g_k^T.$$

This leads us to the discrete K-L expansion of the random vector v :

$$v = \sum_{k=1}^n \sqrt{\lambda_k} \xi_k g_k$$

where $\{\xi_k\}_{k=1, \dots, n}$ is a sequence of independent standard Gaussian random variables.

So what remains here, is to solve the eigenvalue problem (15). If the dimension of B_h is not large, one may use standard numerical methods, e.g., the Lanczos algorithm. However to approximate random fields with high accuracy, one needs to take a subdivision which is fine enough, so the matrix B_v can be of very large size. Then, we can use the randomized low rank approximation method described in section 2.2. It should be noted that the method can be very efficient if the matrix B_v admits a good low rank approximation which is in many practical cases true when the correlation is not too long-ranged.

Lorenzian Random Field. In [34], we have presented the following results of simulation obtained by the randomized SVD based algorithm described.

Let us consider the following example [30] where we have considered the following random boundary value problem: in the upper half-plane $G = \{(x, y) : y \geq 0\}$, find a solution to the Laplace equation $\Delta u(x, y) = 0$ with the boundary conditions $u|_{y=0} = g(x)$ where $g(x)$ is a Gaussian zero mean white noise. We have constructed the solution explicitly, which says that the solution $u(x, y)$ is a partially homogeneous (i.e., homogeneous with respect to the longitudinal coordinate x) Gaussian random field which is uniquely defined by its correlations at two points (x_1, y_1) , (x_2, y_2) , and the correlation function has the following Lorenzian form

$$B(x_1, y_1; x_2, y_2) = \langle u(x_1, y_1) u(x_2, y_2) \rangle = \frac{1}{\pi} \frac{y_1 + y_2}{(y_1 + y_2)^2 + (x_1 - x_2)^2} . \quad (16)$$

Thus the random process $u(x, y)$ is inhomogeneous in transverse direction. In [30], we have found an explicit K-L expansion of this solution, so it was used to validate our randomized SVD based algorithm. The solution $u(x, y)$ on a rectangular G with a grid with 500×500 nodes was simulated, and the rank $k = 20$ approximation was already enough to calculate the solution with 1%-accuracy. The number of randomly sampled rows in the randomized SVD algorithm was $s = 200$. The reason why the rank $k = 20$ was enough is in the relative rapid decrease of the correlations. In the next example we deal with a long-range correlation function of the fractional Wiener process.

Fractional Wiener Process. Let us consider the fractional Wiener process $W^H(t)$ of index H , $H \in (0, 1)$ (Hurst parameter) which is defined as a centered Gaussian inhomogeneous random process on $[0, 1]$ with the following correlation function

$$B_H(s, t) = E[W^H(s)W^H(t)] = \frac{1}{2} \left(s^{2H} + t^{2H} - |t - s|^{2H} \right) .$$

Simulation results for the fractional Wiener process on the interval $[0, 2.5]$ with the Hurst constant $H = 0.3$ are presented in [35], the randomized SVD algorithm with $k = 80$ rank approximation was constructed by sampling 160 random rows, in the 240×240 correlation matrix.

5 Solution of Integral Equations

5.1 Singular Approximations

The low rank approximation can be used to transform the original integral equation to an equivalent integral equation with a new kernel whose properties are better in certain sense. For instance, in [31], Sect. 2.2 we present a singular approximations based method where the norm of the new kernel of the transformed equations is less than 1. This can be achieved by the randomized SVD

with very low rank approximations. Let us present the method for a system of linear algebraic equations, for details of numerical simulation see [35].

Thus we consider a large system of linear equations with an $m \times m$ matrix and right-hand side vector $b = (b_1, \dots, b_m)^T$, and it is assumed that $\|A\| \geq 1$, hence the Neumann series diverges. We introduce a matrix

$$B = A - \sum_{i=1}^r \alpha_i \beta_i^T \quad (17)$$

where $\alpha_1, \dots, \alpha_m$ and $\beta_1, \dots, \beta_m$ are arbitrary column-vectors, i.e., the matrix B is obtained by subtraction from A a sum of singular matrices of the form $\alpha_i \beta_i^T$. Suppose such matrices are found, and we are interested in the relation between the solution x and the solution of the equation with the matrix B .

Consider $r + 1$ auxiliary linear systems with the matrix B for different right-hand sides:

$$x_0 = Bx_0 + b, \quad x_1 = Bx_1 + \alpha_1, \quad \dots \quad x_r = Bx_r + \alpha_r. \quad (18)$$

Then

$$x = x_0 + \sum_{i=1}^r J_i x_i \quad (19)$$

where $J_1, \dots, J_r$ are components of the vector J which satisfies the equation $J = TJ + t$ where T is the matrix with entries $T_{ij} = \beta_i^T x_j$, $i, j = 1, \dots, r$, and t is a vector with components $t_i = \beta_i^T x_0$, $i = 1, \dots, r$.

Practical implementation of this method has a sense if for small value of r we can find the expansion (17) with $q_B = \|B\| < 1$. Note that the randomized SVD algorithm suggests such a solution, and we can try, step by step, to increase the number of terms till the condition $q_B = \|B\| < 1$ is satisfied. For example, in the boundary integral equation formulation of the Laplace equation for a convex domain one may take $r = 1$ (e.g., see [17]). For non-convex domains, r can be chosen quite small, as our calculations presented in the next section show. This is true for quite general singular kernels of the potential theory which appear in the relevant boundary integral equations, see, e.g., [25], [28], [29].

5.2 Inconsistent Systems, Linear Least Squares, and Ill-Posed Problems

The general formulation of a linear least squares problem is the following: we have a set of vectors which we wish to combine linearly to provide the best possible approximation to a given vector. If the set of vectors is $\{a_1, a_2, \dots, a_n\}$ and the given vector is b , we seek coefficients $x_1, x_2, \dots, x_n$ which produce a minimal error $|b - \sum_{i=1}^n x_i a_i|$.

We have to choose the vector x so as to minimize $|Ax - b|$. Let the SVD of A be $U \Sigma V^T$ (where U and V are square orthogonal matrices, and Σ is rectangular with the same dimensions as A). Then we have

$$Ax - b = U \Sigma V^T x - b = U(\Sigma V^T x) - U(U^T b) = U(\Sigma y - c) \quad (20)$$

where $y = V^T x$ and $c = U^T b$. Note that U is an orthogonal matrix, and so preserves lengths, i.e., $|U(\Sigma y - c)| = |\Sigma y - c|$, and hence $|Ax - b| = |\Sigma y - c|$. This suggests a method for solving the least squares problem. First, determine the SVD of A and calculate c as the product of U^T and b . Then, solve the least squares problem for Σ and c , i.e., find a vector y so that $|\Sigma y - c|$ is minimal which is obviously trivial since Σ is diagonal. Now, $y = V^T x$ so we can determine x as Vy . That gives the solution vector x as well as the magnitude of the error, $|\Sigma y - c|$.

6 Random Row Action Iteration Process

We describe here a randomized version of the projection methods belonging to the class of a "row-action" methods which work well both for systems with singular matrices and for overdetermined systems. These methods belong to a type known as *Projection on Convex Sets* methods. Here we present a method beyond the conventional Markov chain based Neumann–Ulam scheme. The main idea is in a random choice of the row in the projection method so that in average, the convergence is improved compared to the conventional periodic choice of the rows. We extend this randomized method for solving linear systems coupled with systems of linear inequalities.

The row action iteration process also known as the projection method suggested first by Kaczmarz [22] can be proved to converge for any system of linear equations with nonzero rows, even when it is singular and inconsistent and the arithmetic operations required in an iteration of the method are comparatively few.

Let us consider a system of linear algebraic equations

$$Ax = b \quad (21)$$

where A is a rectangular $m \times n$ matrix with $m \geq n$, and $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$.

We further denote by a_i the i -th row of A , and a_i^T is the relevant column-vector, the transpose of a_i .

Our stochastic iterative process is written as follows

$$x_{k+1} = x_k + \omega_k E \frac{b_{\nu(i)} - (a_{\nu(i)} \cdot x_k)}{\|a_{\nu(i)}\|^2} a_{\nu(i)}^T, \quad k = 1, 2, \dots \quad (22)$$

where ω_k are some parameters (could be random), the expectation E is taken over the distribution of random indices $\nu(i)$ whose values are sampled at random among random subsets of indices lying in $(1, 2, \dots, m)$. We show that the distribution can be chosen so that the method converges with expected exponential rate, not depending on the number of equations in the system. The solver does not even need to know the whole system, but only some random rows of the matrix, therefore, it is well suited for solving very large systems of linear algebraic equations. Moreover, this method can be used for solving systems of linear equations coupled with systems of linear inequalities. Remarkably, the structure of the algorithm remains practically the same. We note that an example of non-uniform sampling of the random rows in the row action process was suggested

in [39] which is quite costly, because it requires recalculation of the sampling probabilities in each iteration process.

So assume we solve a coupled system of linear equations and inequalities

$$a_i^T x \leq b_i \quad i \in I_{\leq}, \quad (23)$$

$$a_i^T x = b_i \quad i \in I_{=} . \quad (24)$$

Let

$$\gamma_k^{(i)} = \begin{cases} [(a_i \cdot x_k) - b_i]^+ & \text{if } i \in I_{\leq} \\ (a_i \cdot x_k) - b_i & \text{if } i \in I_{=} , \end{cases}$$

and write the iteration process in the form:

$$x_{k+1} = x_k - \frac{\gamma_k^{(\nu(i))}}{\|a_{\nu(i)}\|^2} a_{\nu(i)}^T, \quad k = 1, 2, \dots . \quad (25)$$

It can be shown that this process is convergent, and

$$E[d^2(x_{k+1}, S)] \leq \left(1 - \frac{1}{L^2 \|A\|_F^2}\right) d^2(x_k, S) .$$

Here L is the Hoffmann constant defined by

$$d(x, S_b) \leq L \|e(Ax - b)\|$$

where S_b is the set of possible solutions of our systems, $d(x, S_b)$ is the Euclidean distance from x to the set S_b , and $e(y)$ defines the error in the relevant line of our system of equations and inequalities

$$e(y)_i = \begin{cases} y_i^+ & (i \in I_{\leq}) \\ y_i & (i \in I_{=}) \end{cases}$$

References

1. Achlioptas, D., McSherry, F.: Fast computation of low rank matrix approximations. In: Proceedings of the 33rd Annual Symposium on Theory of Computing (2001)
2. Ailon, N., Chazelle, B.: The fast JohnsonLinderstrauss transform and approximate nearest neighbors. *SIAM J. Comput.* 39(1), 302–322 (2009)
3. Belongie, S., Fowlkes, C., Chung, F., Malik, J.: Spectral Partitioning with Indefinite Kernels Using the Nystrom Extension. In: Heyden, A., et al. (eds.) *ECCV 2002*. LNCS, vol. 2352, pp. 531–542. Springer, Heidelberg (2002)
4. Beylkin, G., Mohlenkam, M.J.: Algorithms for numerical analysis in high dimension. *SIAM Journal on Scientific Computing* 26(6), 2133–2159 (2005)
5. Dimov, I., Philippe, B., Karaivanova, A., Weihrauch, C.: Robustness and Applicability of Markov Chain Monte Carlo Algorithms for Eigenvalue Problem. *Journal of Applied Mathematical Modelling* 32, 1511–1529 (2008)
6. Dimov, I., Alexandrov, V., Papancheva, R., Weihrauch, C.: Monte Carlo Numerical Treatment of Large Linear Algebra Problems. In: Shi, Y., et al. (eds.) *ICCS 2007*. LNCS, vol. 4487, pp. 747–754. Springer, Heidelberg (2007)

7. Dimov, I.T.: Monte Carlo Methods for Applied Scientists, p. 291. World Scientific, Singapore (2008)
8. Drineas, P., Frieze, A., Kannan, R., Vempala, S., Vinay, V.: Clustering Large Graphs via the Singular Value Decomposition. *Machine Learning* 56(13), 9–33 (2004)
9. Drineas, P., Kannan, R.: Pass Efficient Algorithms for Approximating Large Matrices. In: *Proceedings of the 14th Annual Symposium on Discrete Algorithms* (Baltimore, MD), pp. 223–232 (2003)
10. Drineas, P., Drinea, E., Huggins, P.S.: An experimental evaluation of a Monte Carlo algorithm for singular value decomposition. In: Manolopoulos, Y., Evripidou, S., Kakas, A.C. (eds.) *PCI 2001. LNCS*, vol. 2563, pp. 279–296. Springer, Heidelberg (2003) ISSN 0302-9743
11. Drineas, P., Kannan, R.: Fast Monte Carlo algorithms for approximate matrix multiplication. In: *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, p. 452 (2001) ISBN: 0-7695-1390-5
12. Drineas, P., Kannan, R., Mahoney, M.W.: Fast Monte Carlo algorithms for matrices I: approximating matrix multiplication. *SIAM J. Comput.* 36(1), 132–157 (2006)
13. Eberly, W., Kaltfen, E.: On Randomized Lanczos Algorithms. In: *International Conference on Symbolic and Algebraic Computation Archive Proceedings of the 1997 International Symposium on Symbolic and Algebraic Computation*, pp. 176–183 (1997)
14. Eckhart, C., Young, G.: A principal axis transformation for non-Hermitian matrices. *Bulletin of the American Mathematical Society* 45, 118–121 (1939)
15. Ermakov, S.M., Mikhailov, G.A.: Statistical modeling. Nauka, Moscow (1982) (in Russian)
16. Ermakov, S.M.: Monte Carlo Method in Computational Mathematics. An Introductory course. BINOM publisher, St. Petersburg (2009) (in Russian)
17. Ermakov, S.M., Sipin, A.S.: A new Monte Carlo scheme for solving problems of mathematical physics. *Soviet Dokl.* 285(3) (1985) (Russian)
18. Frieze, A., Kannan, R., Vempala, S.: Fast Monte Carlo algorithms for finding low-rank approximations. *J. ACM* 51(6), 1025–1041 (2004)
19. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
20. Hammersley, J.M., Handscomb, D.C.: *Monte Carlo Methods*. Chapman and Hall, London (1964)
21. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math.* 26, 189–206 (1984)
22. Kaczmarz, S.: Angenaeuerte Aufloesung von Systemen linearer Gleichungen. *Bull. Acad. Polon. Sciences et Lettres, A*, 355–357 (1937)
23. Kobayashi, M., Dupret, G., King, O., Samukawa, H.: Estimation of singular values of very large matrices using random sampling. *Computers and Mathematics with Applications* 42, 1331–1352 (2001)
24. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards* 45(4), 255–282 (1950)
25. Liberty, E., Woolfe, F., Martinsson, P.-G., Rokhlin, V., Tygert, M.: Randomized algorithms for the low-rank approximation of matrices. *Yale Dept. of Computer Science Technical Report* 1388
26. Muller, N., Magaia, L., Herbst, B.M.: Singular Value Decomposition, Eigenfaces, and 3D Reconstructions. *SIAM Review* 46(3), 518–545 (2004)

27. Phoon, K.K., Huang, H.W., Quek, S.T.: Simulation of strongly non-Gaussian processes using Karhunen-Loeve expansion. *Probabilistic engineering Mechanics* 20, 188–198 (2005)
28. Rokhlin, V.: Rapid solution of integral equations of classical potential theory. *J. Comp. Phys.* 60, 187–207 (1985)
29. Rokhlin, V., Szlam, A., Tygert, M.: A randomized algorithm for principal component analysis. *SIAM J. Matrix Anal. Appl.*, arxiv.org (2009)
30. Expansion of random boundary excitations for some elliptic PDEs. *Monte Carlo Methods and Applications* 13(5-6), 403–451 (2007)
31. Sabelfeld, K.K.: *Monte Carlo Methods in Boundary Value Problems*. Springer, Heidelberg (1991)
32. Sabelfeld, K.K., Simonov, N.A.: *Random Walks on Boundary for Solving PDEs*. VSP, The Netherlands, Utrecht (1994)
33. Sabelfeld, K., Loshina, N.: Fast stochastic iterative projection methods for very large linear systems. In: *Seventh IMACS Seminar on Monte Carlo Methods (MCM 2009)*, Brussels, September 6-11 (2009)
34. Sabelfeld, K., Mozartova, N.: Sparsified Randomization Algorithms for large systems of linear equations and a new version of the Random Walk on Boundary method. *Monte Carlo Methods and Applications* 15(3), 257–284 (2009)
35. Sabelfeld, K., Mozartova, N.: Sparsified Randomization Algorithms for low rank approximations and applications to integral equations and inhomogeneous random field simulation. *Mathematics and Computers in Simulation* (2010) (submitted)
36. Sabelfeld, K., Shalimova, I., Levykin, A.: Random Walk on Fixed Spheres for Laplace and Lamé equations. *Monte Carlo Methods and Applications* 12(1), 55–93 (2006)
37. Sobol, I.M.: *Numerical Monte Carlo Methods*. Nauka, Moscow (1973) (in Russian)
38. Strang, G.: The fundamental Theorem of linear algebra. *The American Mathematical Monthly* 100(9), 848–855 (1993)
39. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications* 15, 262–278 (2009)
40. Stewart, G.W.: On the Early History of the Singular Value Decomposition. *SIAM Review* 35(4) (1993)
41. Vempala, S.S.: *The Random projection method*. AMS (2004)
42. Vorobiev, Ju.V.: Stochastic iteration process. *J. Comp. Math. and Math. Physics* 4(6), 5(5), 1088–1092, 787-795 (1964) (in Russian)
43. Walker, A.J.: New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronic Letters* 10, 127–128 (1974)
44. Woolfe, F., Liberty, E., Rokhlin, V., Tygert, M.: A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis* 25, 335–366 (2008)

SM Stability for Time-Dependent Problems

Petr N. Vabishchevich

Keldysh Institute of Applied Mathematics, RAS
4 Miusskaya Square, 125047 Moscow, Russia
vabishchevich@gmail.com

Abstract. Various classes of stable finite difference schemes can be constructed to obtain a numerical solution. It is important to select among all stable schemes such a scheme that is optimal in terms of certain additional criteria. In this study, we use a simple boundary value problem for a one-dimensional parabolic equation to discuss the selection of an approximation with respect to time. We consider the pure diffusion equation, the pure convective transport equation and combined convection-diffusion phenomena. Requirements for the unconditionally stable finite difference schemes are formulated that are related to retaining the main features of the differential problem. The concept of SM stable finite difference scheme is introduced. The starting point are difference schemes constructed on the basis of the various Padé approximations.

1 Introduction

When time-dependent problems of mathematical physics are solved numerically, much emphasis is placed on computational algorithms of higher orders of accuracy (e.g., see [1, 2]). Along with improving the approximation accuracy with respect to space, improving the approximation accuracy with respect to time is also of interest. In this respect, the results concerning the numerical methods for ordinary differential equations (ODEs) [3, 4] provide an example. Taking into account the specific features of time-dependent problems for PDEs, we are interested in numerical methods for solving the Cauchy problem in the case of stiff equations [5–7].

When time-dependent problems are solved approximately, the accuracy can be improved in various ways. In the case of two-level schemes (the solution at two adjacent time levels is involved), polynomial approximations of the scheme operators on the solutions are used explicitly or implicitly. The most popular representatives of such schemes are Runge-Kutta methods [7, 8], which are widely used in modern computations. The main feature of the multilevel schemes (multistep methods) manifests itself in the approximation of time derivatives with a higher accuracy on a multipoint stencil. A characteristic example is provided by multistep methods based on backward numerical differentiation [9].

Various classes of stable finite difference schemes can be constructed to obtain a numerical solution [10, 11]. It is important to select among all stable schemes such a scheme that is optimal in terms of certain additional criteria. In the theory

of finite difference schemes, there is the class of asymptotically stable schemes (see [12, 13]) that ensure the correct long-time behavior of the approximate solution. In the theory of numerical methods for ODEs (see [7, 9]), the concept of L -stability is used, which reflects the long-time asymptotic behavior of the approximate solution from a different point of view.

In [14] the properties of two-level difference schemes of high order approximation for the approximate solution of the Cauchy problem for evolutionary equations with self-adjoint operators are considered. The simplest boundary value problem for the one-dimensional parabolic equation serves as a basic problem. The concept of SM stability (Spectral Mimetic stability) of a difference scheme is introduced. This property is connected with the behavior of individual harmonics of the approximate solutions.

In this paper, we continue to study the SM properties of difference schemes for the approximate solutions of unsteady problems of mathematical physics. On the model boundary value problem for one-dimensional parabolic equation, the spectral characteristics of the approximations in space and in time are considered. In particular, good approximation properties (third order approximation in space) are observed for the convection operator. Two-level schemes of higher order of approximation in time, based on the Padé approximation, are considered for solving problems of mathematical physics with symmetric and skew-symmetric operators.

2 Problem Formulation

We consider finite-dimensional real Hilbert space H , where the scalar product and the norm are $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. Let $u(t)$ ($0 \leq t \leq T > 0$) be defined as the solution of the Cauchy problem for evolutionary equation of first order:

$$\frac{du}{dt} + \Lambda u = f(t), \quad 0 < t \leq T, \quad (1)$$

$$u(0) = u_0. \quad (2)$$

The right-hand side $f(t) \in H$ of equation (1) is given and Λ , depending on t ($\Lambda = \Lambda(t) \geq 0$), is a linear non-negative, in generally, not self-adjoint operator from H to H .

For problem (1), (2) the estimate of stability is easily established. Taking into account the skew-symmetric property of operator Λ , we have the equality

$$\|u\| \frac{d\|u\|}{dt} = (f, u).$$

By using

$$(f, u) \leq \|u\| \|f\|$$

we obtain a simple estimate of stability for the solution of (1), (2) with respect to the initial data and the right-hand side:

$$\|u(t)\| \leq \|u_0\| + \int_0^t \|f(\theta)\| d\theta. \quad (3)$$

We would like to preserve these properties of the differential problem after the transition to a discrete analogue of problem (1), (2).

The main attention in our discussion is given to unsteady boundary value problem for partial differential equations. In this context, we can associate the Cauchy problem (1), (2) with the application of the method of lines (approximation in space). Having in mind the importance for applications, we will direct our considerations on an example of a boundary value problem for the one-dimensional parabolic equation of second order. Let a sufficiently smooth function $u(x, t)$ satisfies the equation

$$\frac{\partial u}{\partial t} + \mathcal{L}u = 0, \quad 0 < x < 1, \quad 0 < t \leq T \quad (4)$$

and the initial condition

$$u(x, 0) = u_0(x), \quad 0 < x < 1. \quad (5)$$

Periodicity of the solution with respect to the spatial variable is assumed:

$$u(x + 1, t) = u(x, t), \quad 0 < t \leq T. \quad (6)$$

We associate operator $\mathcal{L}$ with the convection-diffusion equation, defining

$$\mathcal{L} = \chi \mathcal{C} + (1 - \chi) \mathcal{D} \quad (7)$$

for some constant $0 \leq \chi \leq 1$. Here the operators of convective and diffusive transport are defined as follows

$$\mathcal{C}u = \frac{\partial u}{\partial x}, \quad \mathcal{D}u = -\frac{\partial^2 u}{\partial x^2}. \quad (8)$$

If $\chi = 1$ equation (4) is the convection transport equation whereas if $\chi = 0$ it is the diffusion equation.

The discrete problem should inherit the main properties of the differential problem. In model problem (4)–(6) the skew-symmetric property of operator $\mathcal{C}$ as well as the self-adjoint and non-negative properties of operator $\mathcal{D}$ should be preserved

$$\mathcal{C} = -\mathcal{C}^*, \quad \mathcal{D} = \mathcal{D}^* \geq 0 \quad (9)$$

in space $L_2(0, 1)$ for functions satisfying (6). Stability of the solution of the corresponding problem (1), (2) (estimate (3)) is provided by similar properties of the grid analogs of convective and diffusive transport operators. In our research, we are concentrating on the spectral characteristic of the solution (Spectral Mimetic Properties for grid approximations), when considering the behavior of individual harmonics of the approximate solution.

3 SM Properties of the Approximation in Space

Let us introduce a uniform grid with step h :

$$x_i = ih, \quad i = 0, \pm 1, \pm 2, \dots, \quad Mh = 1,$$

$$\omega = \{x_i \mid i = 0, 1, \dots, M-1\}.$$

We use standard index-free notations of the theory of difference schemes [10]. Let $w = w_i = w(x_i)$, and for the left, right and central difference derivatives we set

$$\begin{aligned}\partial_- w &= \frac{w_i - w_{i-1}}{h}, & \partial_+ w &= \frac{w_{i+1} - w_i}{h}, \\ \partial_0 w &= \frac{1}{2}(\partial_- w + \partial_+ w) = \frac{w_{i+1} - w_{i-1}}{2h}\end{aligned}$$

respectively.

After approximation in space we pose the corresponding to (4)–(6) discrete problem

$$\frac{dy}{dt} + \Lambda y = 0, \quad x \in \omega, \quad 0 < t \leq T, \quad (10)$$

$$y(x, 0) = u_0(x), \quad x \in \omega. \quad (11)$$

Some key possibilities in the choice of the grid operator Λ , connected with the properties of the differential operator $\mathcal{L}$, should be noted.

We define Hilbert space $H = L_2(\omega)$ of periodic grid functions ($y(x+1) = y(x)$) with the inner product and the norm

$$(y, w) = \sum_{x \in \omega} y(x) w(x) h, \quad \|y\|^2 = (y, y).$$

To guarantee stability of the solution to problem (10), (11), the operator Λ must be non-negative ($\Lambda \geq 0$) in H . Conservatism (neutral stability) communicates directly with the skew-symmetric property of operator Λ ($\Lambda = -\Lambda^*$) in H .

For the convection equation ($\chi = 1$, $\mathcal{L} = \mathcal{C}$) with $f = 0$ the norm of the solution of problem (4)–(6) does not change in time:

$$\|u(t)\| = \|u_0\|. \quad (12)$$

Equality (12) reflects the conservation property of the solution (conservation law), the neutral stability of the solution.

1. Upwind (directional) approximations of first order. To approximate the convective terms (see, e.g., [1, 15, 16]), the upwind first-order approximations are traditionally widely used. In this case, the grid convection operator C has the form

$$C = \partial_-. \quad (13)$$

Operator C defined in (13) is non-negative ($C \geq 0$). In this case, for the solution of problem (10), (11) the estimate

$$\|y\| \leq \|u_0\|, \quad 0 < t \leq T \quad (14)$$

is true.

2. Central-difference approximation. Another well-known variant is to use approximations of second order, where

$$C = \partial_0. \quad (15)$$

In this case we have $C = -C^*$ and for problem (10), (11), (15) it holds

$$\|y\| = \|u_0\|, \quad 0 < t \leq T. \quad (16)$$

3. Upwind second-order approximations. When choosing approximations of higher order (second and above) for the convective terms, we are trying at least partially to preserve the properties of the first order approximations, which are connected primarily with the monotonicity (fulfillment of the maximum principle). The most interesting attempts in the class of linear approximations are associated with the use of approximations with the upwind differences of second order [2, 17]. For our problem (4)–(6), we have

$$Cy = \frac{3y_i - 4y_{i-1} + y_{i-2}}{2h}.$$

Using previously introduced operator notations we obtain

$$C = \partial_- + \frac{h}{2} \partial_- \partial_-. \quad (17)$$

Operator $C \geq 0$ and so for the solution of problem (10), (11) estimate (14) holds again.

4. Approximations of third order. In computing practice third order approximations are not in common use. In fact, they are only mentioned (see, e.g., [2, 4]) without any meaningful analysis. In this case the difference convection operator can be written in the form

$$C = \partial_0 - \frac{h^2}{6} \partial_- \partial_- \partial_+. \quad (18)$$

In index notation equation (18) takes the form

$$Cy = \frac{2y_{i+1} + 3y_i - 6y_{i-1} + y_{i-2}}{6h}.$$

Operator $C \geq 0$ and its energy (equal to (Cy, y)) is three times less than the energy of operator C , which is defined by rule (17) (upwind second-order approximations).

The stability conditions (neutral stability) of the considered approximations of convection transfer are associated with the general properties of the operator (non-negativity, skew-symmetric property). More detailed information gives us the spectrum of the difference operator, its proximity to the spectrum of the differential operator. This inheritance of the properties of the differential problem by the difference problem at spectral level we associate [14] with the SM properties.

Consider the corresponding differential problem for eigenvalues and eigenfunctions. For operator $\mathcal{C}$ we have

$$\frac{dv}{dx} = \lambda v, \quad 0 < x < 1, \quad (19)$$

$$v(x+1) = v(x). \quad (20)$$

The solution of spectral problem (19),(20) is

$$\lambda_m = i2\pi m, \\ v_m(x) = e^{i2\pi mx}, \quad m = 0, \pm 1, \pm 2, \dots$$

For the solution of problem (4)–(6) we obtain the representation

$$u(x, t) = \sum_{m=-\infty}^{\infty} (u_0, v_m) e^{-\lambda_m t} v_m(x), \quad (21)$$

where

$$(u_0, v_m) = \int_0^1 u_0(x) v_m(x) dx, \quad m = 0, \pm 1, \pm 2, \dots$$

are coefficients of the expansion for function $u_0(x)$.

We now consider the corresponding discrete spectral problems

$$Cv = \mu v \quad (22)$$

with the above-mentioned approximations of the convective term. For definiteness, we assume that M is odd. The eigenfunctions of problem (22) for the difference operators (13), (15), (17) and (18) have the form

$$w_m(x) = e^{i2\pi mx}, \quad x \in \omega, \quad m = 0, \pm 1, \pm 2, \dots, \frac{M-1}{2}. \quad (23)$$

For the difference operator C , defined by (13), the eigenvalues have the form

$$\mu_m = \frac{1 - e^{i2\pi mh}}{h}, \quad m = 0, \pm 1, \pm 2, \dots, \frac{M-1}{2}.$$

For the imaginary and the real parts we obtain

$$\operatorname{Re} \mu_m = \frac{2}{h} \sin^2(\pi mh), \quad \operatorname{Im} \mu_m = \frac{1}{h} \sin(2\pi mh), \quad m = 0, \pm 1, \pm 2, \dots, \frac{M-1}{2}. \quad (24)$$

For the central-difference approximations (15) we obtain

$$\operatorname{Re} \mu_m = 0, \quad \operatorname{Im} \mu_m = \frac{1}{h} \sin(2\pi mh), \quad m = 0, \pm 1, \pm 2, \dots, \frac{M-1}{2}. \quad (25)$$

Comparing (24) and (25) we find that the imaginary components of the spectrum of central-difference approximations and upwind difference approximations coincide. For upwind approximations we have positive real parts of the spectrum, which cause the dissipative properties of such approximations.

Dissipative properties demonstrate also approximations (17), (18). For the upwind second-order approximations we have

$$\operatorname{Re} \mu_m = \frac{1}{h} (\cos(2\pi mh) - 1)^2, \quad \operatorname{Im} \mu_m = \frac{1}{h} \sin(2\pi mh) (2 - \cos(2\pi mh)) \quad (26)$$

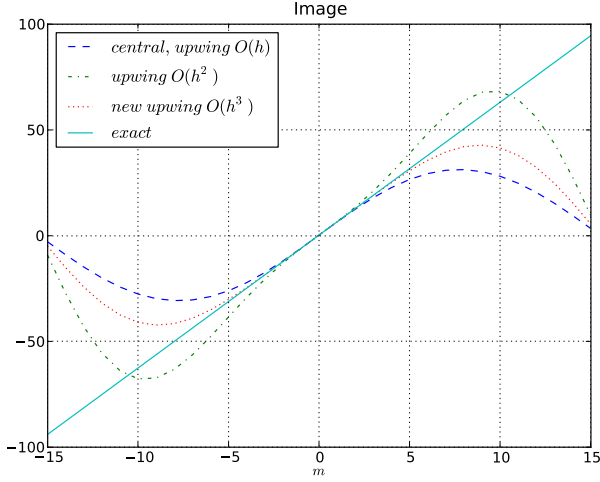


Fig. 1.

with the above-mentioned values of m . For the third order approximations it is easy to obtain

$$\operatorname{Re} \mu_m = \frac{1}{3h} (\cos(2\pi mh) - 1)^2, \quad \operatorname{Im} \mu_m = \frac{1}{3h} \sin(2\pi mh) (4 - \cos(2\pi mh)) \quad (27)$$

respectively.

An illustration of the spectrum (24)–(27) of the grid convection operator is shown in Fig.1,2 for $M = 31$. In particular, the main disadvantage of the scheme with directional differences of first order is associated with substantial dissipation of the low harmonics whereas dissipative properties of schemes (17), (18) are connected primarily with the high harmonics. Approximation of third-order (18) relatively well reflects the spectral properties of the differential problem. Its dissipative properties work only for high harmonics and are weak for the most important low harmonics of the difference solution.

A similar analysis has been performed for the diffusion equation, where $\chi = 0$, $\mathcal{L} = \mathcal{D}$. In this case, we investigated the non-negativity and self-adjointness of the discrete diffusion operator D and its spectral properties.

1. Approximation of second order. The standard approximation at the three-point stencil leads us to

$$D = -\partial_+ \partial_- . \quad (28)$$

2. Approximation of fourth order. At the extended stencil we can use

$$D = -\partial_+ \partial_- + \frac{h^2}{12} \partial_+ \partial_- \partial_+ \partial_- . \quad (29)$$

Using (28), (29), we have $D = D^* \geq 0$ in H .

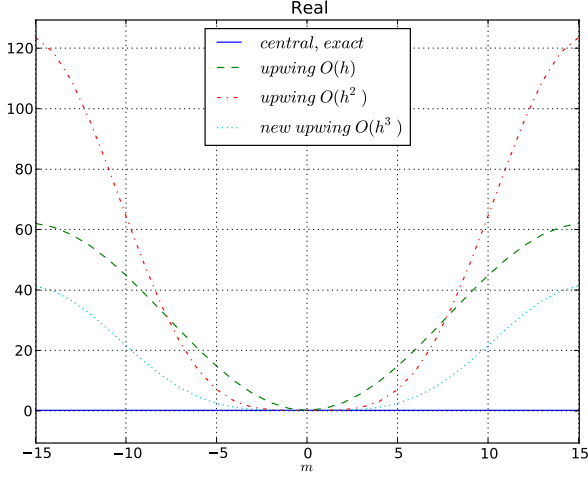


Fig. 2.

The spectrum of these operators D is real with the same eigenfunctions as for C . For the eigenvalues we have

$$\mu_m = \frac{4}{h^2} \sin^2 \frac{m\pi}{M}, \quad m = 0, 1, \dots, M-1$$

for approximation (28) and

$$\mu_m = \frac{4}{h^2} \sin^2 \frac{m\pi}{M} \left(1 + \frac{1}{3} \sin^2 \frac{m\pi}{M} \right), \quad m = 0, 1, \dots, M-1$$

for approximation (29). For the differential operator we have $\lambda_m = 4\pi^2 m^2$, $m = 0, 1, \dots$. As expected, approximation (29) gives better approximations for the spectrum of the differential operator of diffusive transport.

4 SM Properties of the Approximation in Time

We'll use the two-level difference schemes to approximate the solution of (10), (11). Define a uniform grid in time with time-step τ

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau, \quad n = 0, 1, \dots, N, \quad \tau N = T\}$$

and let $y_n = y(t_n)$, $t_n = n\tau$. For the exact solution of problem(10), (11), at transition from time level t_n to new time level t_{n+1} we have

$$y(x, t_{n+1}) = e^{-\Lambda\tau} y(x, t_n) = \sum_{m=-(M-1)/2}^{(M-1)/2} (y(x, t_n), w_m) e^{-\mu_m \tau} w_m(x). \quad (30)$$

Two-level difference scheme for problem (10), (11) is written in the canonical operator-difference form

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad n = 0, 1, \dots \quad (31)$$

with some operators A and B . In the Samarskii theory of stability of operator-difference schemes [10–12] stability conditions in the various norms are formulated in the form of operator inequalities for A, B .

Difference scheme (31) is written as follows

$$y_{n+1} = Sy_n, \quad n = 0, 1, \dots, \quad (32)$$

where

$$S = E - \tau B^{-1}A \quad (33)$$

is the operator of transition from one time level to another level, which, in general, may depend on n .

We restrict ourselves to the simplest difference approximation in time for problem (10), (11), which lead to the transition operator

$$S = s(\tau A), \quad (34)$$

where $s(z)$ is a function of stability [6, 7]. With constraints (34) ($\tau A = (\tau A)(\tau A)$, $B = B(\tau A)$) the stability conditions in Hilbert spaces are easily verified on the basis of the properties of function $s(z)$ only. Let $A \geq \delta E$, then

$$\|s(\tau A)\| \leq \max_{\operatorname{Re} z \geq \delta \tau} |s(z)|,$$

and self-adjointness of the operator A is not assumed.

In the case of (34) for the approximate solution at the new time level we have the representation

$$y(x, t_{n+1}) = \sum_{m=-(M-1)/2}^{(M-1)/2} (y(x, t_n), w_m) s(\mu_m \tau) w_m(x). \quad (35)$$

The quality of the difference approximations in time is estimated by comparing (35) with the representation (30) for the model problem (10), (11). The comparison is performed at the level of behavior of individual harmonics and so we are talking about the SM properties for the approximation in time.

For convection problems ($\chi = 1$, $\mathcal{L} = \mathcal{C}$) the spectrum is purely imaginary, and the solution is neutrally stable. After approximation in space using the above directional differences a typical situation is when the imaginary part of the spectrum is complemented by real part. When choosing approximations in time for the considered problems with skew-symmetric operators, we must monitor the behavior of the main imaginary part of the spectrum. This means that in problem (10), (11)

$$A = A_0 + A_1, \quad A_0 = A_0^* = \frac{1}{2}(A + A^*), \quad A_1 = -A_1^* = \frac{1}{2}(A - A^*),$$

the operator Λ_1 is essential in the sense that $\Lambda_0 y \rightarrow 0$ as $h \rightarrow 0$ for sufficiently smooth y . The supporting real part of the spectrum associated with operator Λ_0 , is generated by the approximations in space and plays a minor role in these problems (it should not lead to instability of the difference solution).

The difference scheme for convection problem (10), (11), in which operator $\Lambda \geq 0$ and its antisymmetric part has the major role, is called SM stable, if the difference scheme is stable and neutrally stable at $\Lambda = -\Lambda^*$.

We will construct two-level difference schemes of higher order of accuracy for time-dependent linear problems on the basis of Padé approximations for the operator (matrix) exponent $e^{-\Lambda\tau}$. For e^{-z} we have

$$e^{-z} = R_{lm}(z) + O(z^{l+m+1}), \quad R_{lm}(z) \equiv \frac{P_{lm}(z)}{Q_{lm}(z)},$$

where $P_{lm}(z)$ and $Q_{lm}(z)$ are polynomials of degree l and m , respectively:

$$P_{lm}(z) = \frac{l!}{(l+m)!} \sum_{k=0}^l \frac{(l+m-k)!}{k!(l-k)!} (-z)^k,$$

$$Q_{lm}(z) = \frac{m!}{(l+m)!} \sum_{k=0}^m \frac{(l+m-k)!}{k!(m-k)!} z^k.$$

For equation (10) the application of Padé approximations corresponds to the two-level scheme

$$Q_{lm}(\tau\Lambda) \frac{y_{n+1} - y_n}{\tau} + \frac{1}{\tau} (Q_{lm}(\tau\Lambda) - P_{lm}(\tau\Lambda)) y_n = 0, \quad n = 0, 1, \dots \quad (36)$$

In canonical form (31) difference scheme (36) corresponds to the choice

$$A = \frac{1}{\tau} (Q_{lm}(\tau\Lambda) - P_{lm}(\tau\Lambda)), \quad B = Q_{lm}(\tau\Lambda). \quad (37)$$

The difference schemes for problem (10), (11) with $\Lambda \geq 0$, constructed on the basis of Padé approximations, are stable (absolutely stable) (estimate $\|y_{n+1}\| \leq \|y_n\|$ holds) for $l \leq m$ [6, 7]. It is only necessary among such schemes to highlight the SM-stable schemes.

In the simplest case $m = 1$ we have

$$R_{01}(z) = \frac{1}{1+z} = e^{-z} + O(z^2),$$

$$R_{11}(z) = \frac{1 - \frac{1}{2}z}{1 + \frac{1}{2}z} = e^{-z} + O(z^3).$$

Approximations $R_{01}(z)$ correspond to application of the purely implicit scheme

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda y_{n+1} = 0, \quad n = 0, 1, \dots \quad (38)$$

for approximation of problem (10), (11).

The application of the symmetric scheme (Crank-Nicholson)

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda \frac{y_{n+1} + y_n}{2} = 0, \quad n = 0, 1, \dots \quad (39)$$

corresponds to the choice of approximation $R_{11}(z)$.

The condition of neutral stability of $\|y_{n+1}\| = \|y_n\|$ for two-level scheme (32) will be satisfied at $\|S\| = 1$. Taking into account (34) and for $\Lambda = -\Lambda^*$ this corresponds to the case

$$|s(z)| = |R_{lm}(z)| = 1, \quad \operatorname{Re} z = 0. \quad (40)$$

For purely implicit scheme (38) we have

$$|R_{01}(z)| = \frac{1}{\sqrt{1+y^2}}, \quad z = iy.$$

Thus, the condition of neutral stability is not satisfied — the purely implicit scheme is not SM stable for problems with the main skew-symmetric operator. While for symmetric scheme (39) we obtain

$$|R_{11}(z)| = 1, \quad z = iy.$$

Thus, this scheme is SM stable for the investigated class of problems.

We can make similar conclusions for schemes with Padé approximations at $m > 1$. Only a scheme that is based on approximation R_{mm} is SM stable for problems with the main skew-symmetric operator. When using Padé approximations with $l < m$ the scheme demonstrates dissipative properties due to the approximation in time. Only at $l = m$ the corresponding scheme is neutrally stable.

A similar analysis is carried out (see[14]) for the diffusion equation. In problem (10), (11) with $\Lambda = \Lambda^* \geq 0$ the amplitudes of harmonics with higher numbers damp more quickly in comparison with the amplitudes of harmonics with lower numbers (spectral monotonicity) and damp to zero as $t \rightarrow \infty$ (asymptotic stability). We associate such a behavior of the approximate solutions with the SM properties of approximation in time for the solution of problems with self-adjoint operators.

We say that the difference scheme for problem (10), (11) with $\Lambda = \Lambda^* \geq 0$ is SM stable if it is spectrally monotonic and asymptotically stable.

Difference schemes based on the Padé approximation R_{lm} are SM stable at $l = 0$. Purely implicit scheme (38) belongs to this class of schemes, whereas the symmetric scheme is conditionally SM stable.

The main conclusion of the study is that to approximate the solutions of problems with skew-symmetric operators we must use approximations in time, based on Padé approximations $R_{mm}(z)$. For problems with self-adjoint operators the use of Padé approximation $R_{0m}(z)$ is more preferable. For problems with general not self-adjoint operators, approximation in time can be constructed via decomposition into self-adjoint and skew-symmetric components and further constructing different approximations for them, based on special splitting schemes.

References

1. Hundsdorfer, W., Verwer, J.: Numerical Solution of Time-Dependent Advection-diffusion-reaction Equations. Springer, Berlin (2003)
2. Gustafsson, B.: High Order Difference Methods for Time Dependent PDE. Springer, Berlin (2008)
3. Ascher, U.M.: Numerical Methods for Evolutionary Differential Equations. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008)
4. LeVeque, R.J.: Finite Difference Methods for Ordinary and Partial Differential Equations. Steady-state and Time-dependent Problems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2007)
5. Rakitskii, Y.V., Ustinov, S.M., Chernorutskii, I.G.: Numerical Methods for Solving Stiff Systems. Nauka, Moscow (1979) (in Russian)
6. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II: Stiff and Differential-Algebraic Problems. Springer, Berlin (1996)
7. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley, Hoboken (2008)
8. Dekker, K., Verwer, J.: Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations. North-Holland, Amsterdam (1984)
9. Gear, C.W.: Numerical Initial Value Problems in Ordinary Differential Equations. Prentice-Hall, Englewood Cliffs (1971)
10. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker Inc., New York (2001)
11. Samarskii, A.A., Matus, P.P., Vabishchevich, P.N.: Difference Schemes with Operator Factors. Kluwer Academic Publishers, Dordrecht Hardbound (2002)
12. Samarskii, A.A., Gulin, A.V.: Stability of Difference Schemes. Nauka, Moscow (1973) (in Russian)
13. Samarskii, A.A., Vabishchevich, P.N.: Computational Heat Transfer. Mathematical Modelling, vol. 1. Wiley, Chichester (1995)
14. Vabishchevich, P.N.: Two-Level Finite Difference Scheme of Improved Accuracy Order for Time-Dependent Problems of Mathematical Physics. Computational Mathematics and Mathematical Physics 50(1), 112–123 (2010)
15. Samarskii, A.A., Vabishchevich, P.N.: Methods for Convection-Diffusion Problems. URSS, Moscow (2004) (in Russian)
16. Morton, K.W.: Numerical Solution of Convection-Diffusion Problems. Chapman & Hall, London (1996)
17. Hirsch, C.: Numerical Computation of Internal and External Flows. Fundamentals of Computational Fluid Dynamics. Butterworth-Heinemann, Amsterdam (2007)

Advanced Monte Carlo Techniques in the Simulation of CMOS Devices and Circuits

Asen Asenov

Device Modelling Group, University of Glasgow
asenov@elec.gla.ac.uk

1 Introduction

The years of happy scaling are over and the fundamental challenges that the semiconductor industry faces at technology and device level will deeply affect the design of the next generations of integrated circuits and systems. The progressive scaling of CMOS transistors to achieve faster devices and higher circuit density has fuelled the phenomenal success of the semiconductor industry captured by Moores famous law [1]. Silicon technology has entered the nano CMOS era with 35nm MOSFETs in mass production in the 45nm technology generation. However, it is widely recognised that the increasing variability in the device characteristics is among the major challenges to scaling and integration for the present and next generation of nano CMOS transistors and circuits. Variability of transistor characteristics has become a major concern associated with CMOS transistors scaling and integration [2], [3]. It already critically affects SRAM scaling [4], and introduces leakage and timing issues in digital logic circuits [5]. The variability is the main factor restricting the scaling of the supply voltage, which for the last four technology generations has remained virtually constant, adding to the looming power crisis.

In this paper we describe advanced Monte Carlo simulation techniques that are used to study statistical variability in contemporary and future CMOS technology generations at the levels of physical transistor simulation, compact model and circuit simulation. First we review the major sources of statistical variability in nano CMOS transistors focusing at the 45nm technology generation and beyond and introduce the advanced 3D statistical physical statistical simulation technology and tools used to forecasts the magnitude of statistical variability. Statistical compact models used to transfer the variability information obtained from the physical simulations into the circuit simulation and design domain are discussed next. Sensitivity analysis allows the selection of optimal statistical compact model sets of parameters. The use of statistical compact models is illustrated in the simulation of SRAM cells.

2 Sources of Statistical Variability

The statistical variability in modern CMOS transistors is introduced by the inevitable discreteness of charge and matter, the atomic scale non-uniformity of the interfaces and the granularity of the materials used in the fabrication of

integrated circuits. The granularity introduces significant variability when the characteristic size of the grains and irregularities become comparable to the transistor dimensions. For conventional bulk MOSFETs, which are still the workhorse of the CMOS technology, Random Discrete Dopants (RDD) are the main source of statistical variability [6]. Random dopants are introduced predominantly by ion implantation and redistributed during high temperature annealing. Fig. 1 illustrates the dopant distribution obtained by the atomistic process simulator DADOS by Synopsys. Apart from special correlation in the dopant distribution imposed by the silicon crystal lattice, there may be also correlations introduced by the Coulomb interactions during the diffusion process. Line Edge Roughness (LER) illustrated in Fig. 2 stems from the molecular structure of the photoresist and the corpuscular nature of light. The polymer chemistry of the 193nm lithography used now for few technology generations mainly determines the current LER limit of approx 5nm [7]. In transistors with poly-silicon gate Poly Gate Granularity (PGG) illustrated in Fig. 3 is another important source of variability. This is associated surface potential pinning at the grain boundaries complimented by doping non-uniformity due to rapid diffusion along the grain boundaries [8].

The introduction of high-k/metal gate technology improves the RDD induced variability, which is inversely proportional to the equivalent oxide thickness (EOT). This is due to the elimination of the polysilicon depletion region and

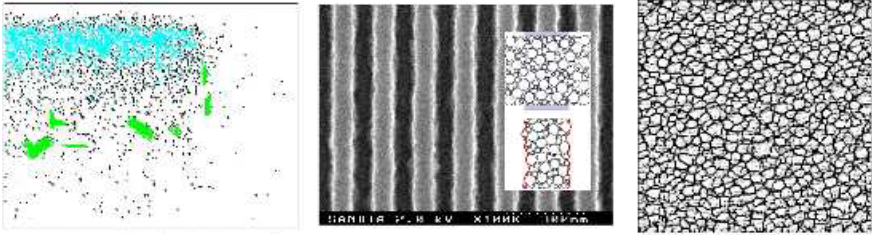


Fig. 1. KMC simulation of RDD (DADOS, Synopsys) **Fig. 2.** Typical LER in Photoresist (Sandia Labs.) **Fig. 3.** SEM micrograph of typical PSG from bottom

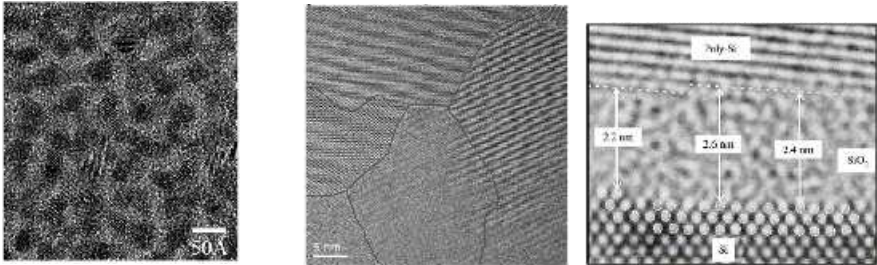


Fig. 4. Granularity in HfON high-k dielectrics (Sematech) **Fig. 5.** Metal granularity causing gate work-function variation **Fig. 6.** Interface roughness (IBM)

better screening of the RDD induced potential fluctuations in the channel from the very high concentration of mobile carriers in the gate. The metal gate also eliminates the PGG induced variability. In the same time it introduces high-k granularity illustrated in Fig. 4 and variability due to work-function variation associated with the metal gate granularity illustrated in Fig. 5 [9]. In extremely scaled transistors atomic scale channel interface roughness illustrated in Fig. 6 [10] and corresponding oxide thickness and body thickness variations [11] can become important source of statistical variability.

3 Simulation of Statistical Variability

The simulation results presented in this chapter were obtained using the Glasgow statistical 3D device simulator, which solves the carrier transport equations in the drift-diffusion approximation with Density Gradient (DG) quantum corrections [22]. In the simulations, the RDD are generated from continuous doping profile by placing dopant atoms on silicon lattice sites within the device S/D and channel regions with a probability determined by the local ratio between dopant and silicon atom concentration. Since the basis of the silicon lattice is 0.543nm a fine mesh of 0.5nm is used to ensure a high resolution of dopant atoms. However, without considering quantum mechanical confinement in the potential well, in classical simulation, such fine mesh leads to carrier trapping at the sharply resolved Coulomb potential wells generated by the ionised discrete random dopants. In order to remove this artifact, the DG approach is employed as a quantum correction technology for both electrons and holes [12].

The LER illustrated in Fig. 2 is introduced through 1D Fourier synthesis. Random gate edges are generated from a power spectrum corresponding to a Gaussian autocorrelation function [7], with typical correlation length $\Lambda=30\text{nm}$ and root-mean-square amplitude $\Delta=1.3\text{nm}$, which is the level that is achieved with current lithography systems [13]. The quoted in the literature values of LER are equal to 3. The procedure used for simulating PGG involves the random generation of poly-grains for the whole gate region [8]: a large atomic force microscope image of polycrystalline silicon grains illustrated at the top of Fig. 3

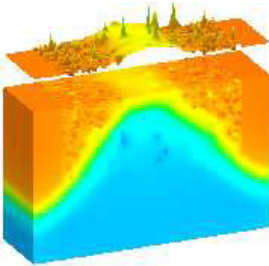


Fig. 7. Potential distribution in a 35nm MOSFET subject to RDD

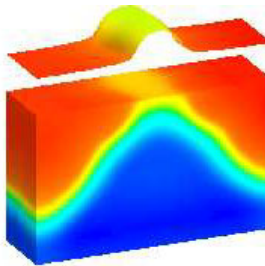


Fig. 8. Potential distribution in a 35nm MOSFET subject to LER

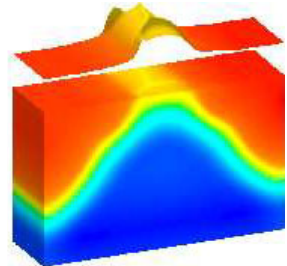


Fig. 9. Potential distribution in a 35nm MOSFET subject to PSG

has been used as a template and the image is scaled according to the average grain diameter experimentally. Then the simulator imports a random section of the grain template image that corresponds to the gate dimension of the simulated device, and along grain boundaries, the applied gate potential in the polysilicon is modified in a way that the Fermi level remains pinned at a certain position in the silicon bandgap. In the worst case scenario the Fermi level is pinned in the middle of the silicon gap. The impact of polysilicon grain boundary variation on device characteristics is simulated through the pinning of the potential in the polysilicon gate along the grain boundaries. The individual impact of RDD, LER and PSG on the potential distribution in a typical 35nm bulk MOSFET is illustrated in Figs. 7, 8 and 9 respectively.

4 Variability in Future Technology Generations

In order to foresee the expected magnitude of statistical variability in the future we have studied the individual impact of RDD, LER and PSG on MOSFETs with gate lengths 35nm, 25nm, 18nm, 13nm and 9nm physical gate length. We also compare the results with the statistical variability introduced in the same devices when RDD, LER and PSG are introduced in the same devices simultaneously. The scaling of the simulated devices is based on a 35nm MOSFET published by Toshiba [14] against which our simulations were carefully calibrated. The scaling closely follows the prescriptions of the ITRS [15] in terms of equivalent oxide thickness, junction depth, doping and supply voltage. The intention was also to preserve the main features of the reference 35nm MOSFET and, in particular, to keep the channel doping concentration at the interface as low as possible. Fig. 10 shows the structure of the scaled devices. More details about the scaling approach and the characteristics of the scaled devices may be found in [12].

Fig. 11 compares the channel length dependence of σV_T introduced by random dopants, line edge roughness and poly-Si grain boundaries with Fermi level pinning. The average size of the polysilicon grains was kept at 40nm for all

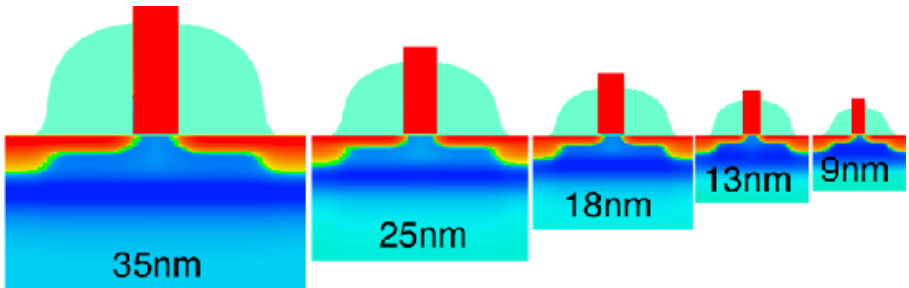


Fig. 10. Examples of realistic conventional MOSFETs scaled from a template 35nm device according to the ITRS requirements for the 90nm, 65nm, 45nm, 32nm and 22nm technologies, obtained from process simulation with Taurus Process

channel lengths. Two scenarios for the magnitude of LER were considered in the simulations. In the first scenario the LER values decrease with the reduction of the channel length following the prescriptions of the ITRS 2003 of 1.2, 1.0, 0.75, and 0.5nm for the 35-, 25-, 18-, and 13-nm channel length transistors, respectively. In this case the dominant source of variability at all channel lengths are the random discrete dopants. The variability introduced by the polysilicon granularity is similar to that introduced by random discrete dopants for the 35nm and 25nm MOSFETs, but at shorter channel lengths the random dopants take over. The combined effect of the three sources of variability is also shown in the same figure. In the second scenario LER remains constant and equal to its current value of approximately 4nm ($\Delta = 1.33$ nm). The results for the 35nm and the 25nm MOSFETs are very similar to the results with scaled LER but below 25nm channel length LER rapidly becomes the dominant source of variability.

Fig. 12 is analogous to Fig. 11 exploring the scenario of the oxide thickness, which is difficult to scale further. The LER is scaled according to the ITRS requirements listed above. Even with the introduction of high-k gate stack it is likely to remain stagnated at 1 nm. This will lead to an explosion in the threshold voltage variability for bulk MOSFETs with physical channel length below 25 nm.

5 Compact Model Strategies for Statistical Variability

It is very important to be able to capture the simulated or measured statistical variability in statistical compact models since this is the only way to communicate this information to designers [16]. Previous research on statistical compact model identification was focused mainly on variability associated with traditional process variations resulting from poor control of critical dimensions, layer

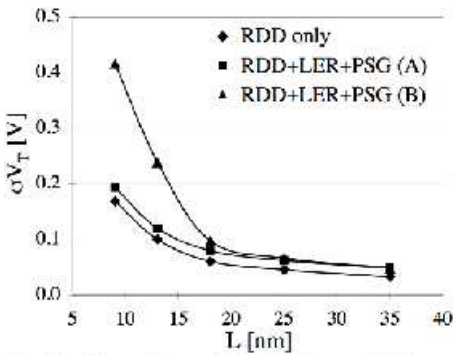


Fig. 11. Channel length dependence of σV_T introduced by random dopants, line edge roughness and poly-Si granularity: (A) LER scales according ITRS; (B) LER= 4nm

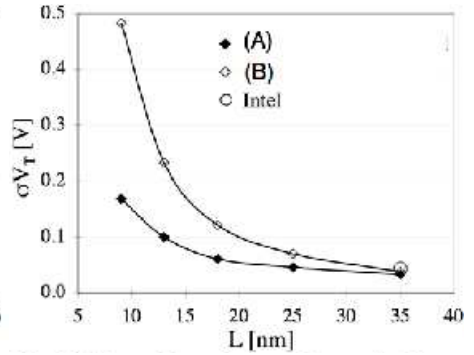


Fig. 12. Channel length dependence of σV_T introduced by random dopants, line edge roughness and poly-Si granularity: (A) t_{ox} scales according ITRS; (B) t_{ox} = 1nm

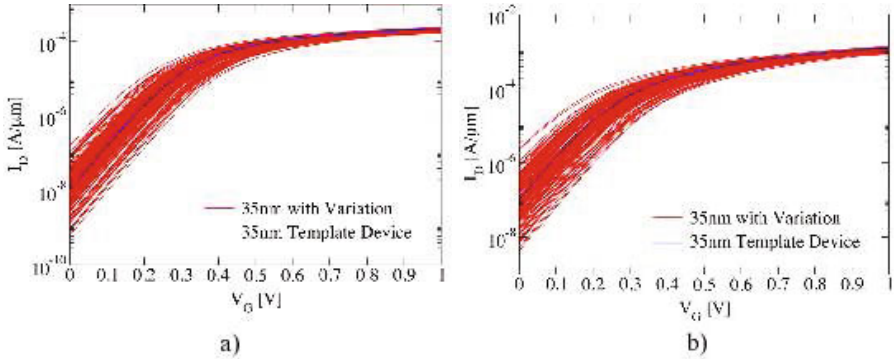


Fig. 13. Variability in the current voltage characteristics of a statistical sample of 200 microscopically different 25nm square (W-L) n-channel MOSFETS at a) $V_D=50$ mV and b) $V_D=1$ V

thicknesses and doping clearly related to specific compact model parameters [17], [18].

Unfortunately, the current industrial strength compact models do not have natural parameters designed to incorporate seamlessly the truly statistical variability associated with RDD, LER, PGG and other relevant variability sources. Despite some attempts to identify and extract statistical compact model parameters suitable for capturing statistical variability introduced by discreteness of charge and matter this remains an area of active research [19]. Fig. 13 shows the spread in $I_D - V_G$ characteristics obtained from atomistic simulator due to the combined effect of RDD, LER and PGG.

We use the standard BSIM4 compact models [20] to capture the information for statistical variability obtained from full 3D physical variability simulation. The statistical extraction of compact model parameters is done in two stages [21]. In the first stage, one complete set of BSIM4 parameters is extracted from the $I - V$ characteristics of uniform (continuously doped, no RDD, LER and PGG) set of devices with different channel lengths and widths and process flow identical to the one of the 35nm testbed transistor [12]. Target current voltage characteristics are simulated over the complete device operating range and parameter extraction strategy combining group extraction and local optimization is employed.

At the second stage, we re-extracted a small carefully chosen subset of the BSIM4 model parameters from the physically simulated characteristics of each microscopically different device in the statistical ensemble keeping the bulk of the BSIM parameters unchanged. The transfer ($I_D - V_G$) characteristics at low and high drain bias are used as extraction target at this stage. The seven re-extracted model parameters are L_{pe0} , R_{dsmin} , N_{factor} , V_{off} , A_1 , A_2 and D_{sub} .

The scatter plots in Fig. 14 show that the chosen seven BSIM4 parameters are not all statistically independent and therefore cannot be generated independently. Also the distributions of most of the seven parameters is not necessarily a normal distribution, which means that it would be difficult to generate

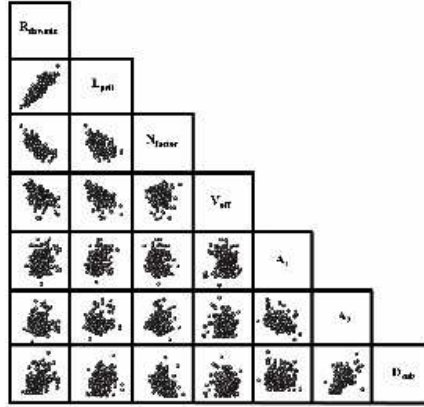


Fig. 14. Scatter plots between each pair of the statistical compact model parameters

statistically accurate parameter sets by employing ordinary multivariate statistical methods, such as principal component analysis. In order to precisely reproduce each individual simulated device, a statistical compact model library is built and statistical instances of devices in statistical circuit simulation can be randomly selected from the library. Statistical weight can be also assigned to each of the statistical compact models for the purposes of statistical enhancement of the statistical circuit simulation.

6 Basics of Statistical Circuit Simulations in the Presence of Statistical Variability

The statistical circuit simulation methodology described in the previous section, which can transfer all the fluctuation information obtained from 3D statistical device simulations into circuit simulation, is employed to investigate the impact of RDF on 6T and 8T SRAM stability for the next three generations of bulk CMOS technology. In the following discussions, we use 25nm, 18nm and 13nm channel length transistors described in details in [22].

Currently, 6T SRAM is the dominant SRAM cell architecture in SoC and microprocessors. However, the disturbance of bit lines on the data retention element during read access makes the 6T cell structure vulnerable to statistical variability, which in turn will have a huge impact on 6T SRAMs scalability. The functionality of SRAM is determined by both static noise margin (SNM) defined as the minimum dc voltage necessary to flip the state of the cell and the write noise margin (WNM) defined as the DC noise voltage needed to fail to flip a cell during a write period. The meaning of SNM and WNM is defined in Fig. 15. Fig. 16 illustrates the statistical nature of SNM and WNM in the presence of statistical transistor variability. The magnitude of WNM in SRAM is mainly determined by the load and access transistors. Since they are the smallest transistors in an SRAM cell, the WNM variation will be larger than the

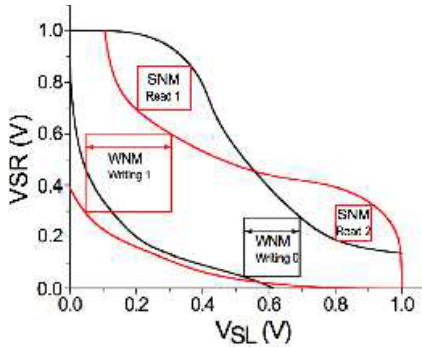


Fig. 15. Static voltage transfer characteristics and definition of SNM and WNM

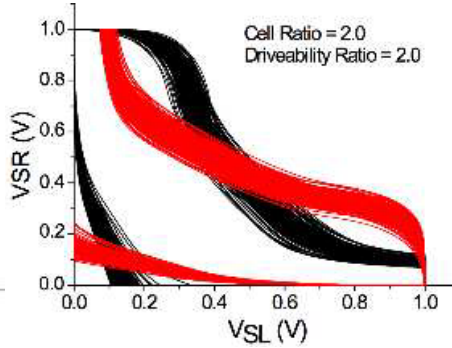


Fig. 16. Statistical behavior of the SNM and WNM of SRAM made of 25 nm bulk MOSFETs subject to RDD^1

SNM variation. However, the mean value of WNM is much larger than its SNM counterpart. Previous studies [23] suggested that under normal circumstances the limiting factor for the operation of bulk 6T SRAM cells is SNM.

7 Conclusions

The statistical variability introduced by discreteness of charge and matter has become one of the major concerns for the semiconductor industry. More and more the strategic technology decisions that the industry will be making in the future will be motivated by the desire to reduce statistical variability. Significant efforts will be needed in the future to improve the statistical aspect of the physical, compact model and circuit level statistical simulations.

References

1. Moore, G.E.: Progress in Digital Electronics. In: Technical Digest of the Int'l Electron Devices Meeting, p. 13. IEEE Press, Los Alamitos (1975)
2. Bernstein, K., Frank, D.J., Gattiker, A.E., Haensch, W., Ji, B.L., Nassif, S.R., Nowak, E.J., Pearson, D.J., Rohrer, N.J.: IBM J. Research and Development 50, 433 (2006)
3. Brown, A.R., Roy, G., Asenov, A.: Poly-Si gate related variability in decanometre MOSFETs with conventional architecture. IEEE Trans. Electron Devices 54, 3056 (2007)
4. Cheng, B.-J., Roy, S., Asenov, A.: The impact of random dopant effects on SRAM cells. In: Proc. 30th European Solid-State Circuits Conference (ESSCIRC), Leuven, p. 219 (2004)
5. Agarwal, A., Chopra, K., Zolotov, V., Blaauw, D.: Circuit optimization using statistical static timing analysis. In: Proc. 42nd Design Automation Conference, Anaheim, p. 321 (2005)

6. Asenov, A.: Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 micron MOSFETs: A 3D atomistic simulation study. *IEEE Trans. Electron Dev.* 45, 2505 (1998)
7. Asenov, A., Kaya, S., Brown, A.R.: Intrinsic Parameter Fluctuations in Decanometre MOSFETs Introduced by Gate Line Edge Roughness. *IEEE Trans. Electron Dev.* 50, 1254 (2003)
8. Brown, A.R., Roy, G., Asenov, A.: Poly-Si gate related variability in decanometre MOSFETs with conventional architecture. *IEEE Trans. Electron Devices* 54, 3056 (2007)
9. Watling, J.R., Brown, A.R., Ferrari, G., Babiker, J.R., Bersuker, G., Zeitzoff, P., Asenov, A.: Impact of High-k on transport and variability in nano-CMOS devices. *J. Computational and Theoretical Nanoscience* 5(6), 1072 (2008)
10. Asenov, A., Kaya, S., Davies, J.H.: Intrinsic Threshold Voltage Fluctuations in Decanano MOSFETs due to Local Oxide Thickness Variations. *IEEE Trans. Electron Dev.* 49, 112 (2002)
11. Brown, A.R., Watling, J.R., Asenov, A.: A 3-D Atomistic Study of Archetypal Double Gate MOSFET Structures. *J. Computational Electronics* 1, 165 (2002)
12. Roy, G., Brown, A.R., Adamu-Lema, F., Roy, S., Asenov, A.: Simulation Study of Individual and Combined Sources of Intrinsic Parameter Fluctuations in Conventional Nano-MOSFETs. *IEEE Trans. Electron Dev.* 52, 3063–3070 (2006)
13. Taiault, J., Foucher, J., Tortai, J.H., Jubert, O., Landis, S., Pauliac, S.: Line edge roughness characterization with three-dimensional atomic force microscope: Transfer during gate patterning process. *J. Vac. Sci. Technol. B* 23, 3070 (2005)
14. Inaba, S., Okano, K., Matsuda, S., Fujiwara, M., Hokazono, A., Adachi, K., Ohuchi, K., Suto, H., Fukui, H., Shimizu, T., Mori, S., Oguma, H., Murakoshi, A., Itani, T., Iinuma, T., Kudo, T., Shibata, H., Taniguchi, S., Takayanagi, M., Azuma, A., Oyamatsu, H., Suguro, K., Katsumata, Y., Toyoshima, Y., Ishiuchi, H.: High performance 35nm gate length CMOS with NO oxynitride gate dielectric and Ni salicide. *IEEE Trans. Electron Dev.* 49, 2263 (2002)
15. International Roadmap for Semiconductors (ITRS), Semiconductor Industry Association, <http://www.itrs.net>
16. Lin, C.-H., Dunga, M.V., Lu, D., Niknejad, A.M., Hu, C.: Statistical Compact Modeling of Variations in Nano MOSFETs VLSI-TSA 2008, p. 165 (2008)
17. Power, J.A., Donellan, B., Mathewson, A., Lane, W.A.: Relating statistical MOSFET model parameter variabilities to ICmanufacturing process fluctuations enabling realistic worst case design. *IEEE Trans. Semicond. Manufact.* 7, 306 (1994)
18. McAndrew, C.C.: Efficient statistical modeling for circuit simulation Design of Systems on a Chip: Devices and Components. In: Reis, R., Jess, J. (eds.), p. 97. Kluwer Academic, Dordrecht (2004)
19. Takeuchi, K., Hane, M.: Statistical Compact Model Parameter Extraction by Direct Fitting to. *IEEE Trans. Electron Devices* 55, 1487 (2008)
20. BSIM4 manual, <http://www-device.eecs.berkeley.edu/>
21. Cheng, B., Roy, S., Roy, G., Adamu-Lema, F., Asenov, A.: Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells. *Solid-State Electron.* 49, 740 (2005)
22. Cheng, B., Roy, S., Asenov, A.: Low power, high density CMOS 6-T SRAM cell design subject to atomistic fluctuations. In: *Proc. ULIS 2006*, Grenoble, pp. 33–36 (2006) ISBN: 88-900874-0-8
23. Cheng, B., Roy, S., Asenov, A.: CMOS 6-T SRAM cell design subject to atomistic fluctuations. *Solid-State Electronics* 51, 565 (2007)

Monte Carlo Method for Numerical Integration Based on Sobol's Sequences

Ivan Dimov and Rayna Georgieva

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
ivdimov@bas.bg, rayna@parallel.bas.bg
<http://parallel.bas.bg/dpa/BG/dimov/index.html>,
<http://parallel.bas.bg/~rayna>

Abstract. An efficient Monte Carlo method for multidimensional integration is proposed and studied. The method is based on Sobol's sequences. Each random point in s -dimensional domain of integration is generated in the following way. A Sobol's vector of dimension s ($\mathcal{M}_\tau$ -point) is considered as a centrum of a sphere with a radius ρ . Then a random point uniformly distributed on the sphere is taken and a random variable is defined as a value of the integrand at that random point. It is proven that the mathematical expectation of the random variable is equal to the desired multidimensional integral. This fact is used to define a Monte Carlo algorithm with a low variance.

Numerical experiments are performed in order to study the quality of the algorithm depending of the radius ρ and regularity, i.e. smoothness of the integrand.

Keywords: Monte Carlo method, multidimensional integration, Sobol's sequences.

1 Introduction and Basic Notation

In this paper we consider numerical algorithms for evaluating multi-dimensional integrals. The problem of evaluating integrals of high dimension is an important task since it appears in many important scientific applications of financial mathematics, economics, environmental mathematics and statistical physics. Randomized (Monte Carlo) algorithms have proven to be very efficient in solving multidimensional integrals in composite domains [2,10].

One may consider two classes of algorithms: *deterministic algorithms* $\mathcal{A}$ and *randomized (Monte Carlo) algorithms* $\mathcal{A}^{\mathcal{R}}$. Usually randomized algorithms reduce problems to the approximate calculation of mathematical expectations. We use the following notation. The mathematical expectation of the random variable (r.v.) θ is denoted by $E_\mu(\theta)$ (sometimes abbreviated to $E\theta$). By $x = (x_1, \dots, x_s)$ we denote a point in a closed domain $\Omega \subset \mathbb{R}^s$, where $\mathbb{R}^s$ is s -dimensional Euclidean space. The s -dimensional unit cube is denoted by $E^s = [0, 1]^s$. We shall further denote the values (realizations) of a random point ξ or r.v. θ by $\xi^{(i)}$ and

$\theta^{(i)} (i = 1, 2, \dots, n)$ respectively. If $\xi^{(i)}$ is a s -dimensional random point, then usually it is constructed using s random numbers γ , i.e., $\xi^{(i)} \equiv (\gamma_{i,1}, \dots, \gamma_{i,s})$. Let I be the desired value of the integral. Assume for a given r.v. θ one can prove that $E\theta = I$. Suppose the mean value of n realizations of θ : $\theta^{(i)}$, $i = 1, \dots, n$ is considered as a Monte Carlo approximation to the solution:

$$\bar{\theta}_n = 1/n \sum_{i=1}^n \theta^{(i)} \approx I. \quad (1)$$

One can only state that a certain randomized algorithm can produce the result with a given probability error.

Definition 1. *If I is the exact solution of the problem, then the probability error is the least possible real number R_n , for which $P = \Pr\{|\bar{\theta}_n - I| \leq R_n\}$, where $0 < P < 1$.*

So, dealing with randomized algorithms one has to accept that the result of the computation can be true only with a certain (even high) probability. In most cases of practical computations it is reasonable to accept an error estimate with a probability smaller than 1.

2 Problem Setting

Consider the following problem of integration:

$$S(f) := I = \int_{E^s} f(x) dx, \quad (2)$$

where $x \equiv (x_1, \dots, x_s) \in E^s \subset \mathbb{R}^s$ and $f \in C(E^s)$ is an integrable function on E^s . The computational problem can be considered as a mapping of function $f : \{[0, 1]^s \rightarrow \mathbb{R}\}$ to $\mathbb{R}$: $S(f) : f \rightarrow \mathbb{R}$, where $S(f) = \int_{E^s} f(x) dx$ and $f \in F_0 \subset C(E^s)$. We will call S the solution operator. The elements of F_0 are the data, for which the problem has to be solved; and for $f \in F_0$, $S(f)$ is the exact solution. For a given f we want to compute (or approximate) $S(f)$.

We will call a *quadrature formula* any expression

$$A = \sum_{i=1}^n c_i f(x^{(i)}),$$

which approximates the value of the integral $S(f)$. The real numbers $c_i \in \mathbb{R}$ are called weights and s dimensional points $x^{(i)} \in E^s$ are called nodes. It is clear that for fixed weights c_i and nodes $x^{(i)} \equiv (x_{i,1}, \dots, x_{i,s})$ the quadrature formula A may be used to define an algorithm. We call a *randomized quadrature formula* any formula of the following kind:

$$A^R = \sum_{i=1}^n \sigma_i f(\xi^{(i)}),$$

where σ_i and $\xi^{(i)}$ are random weights and nodes respectively. The algorithm A^R belongs to the class of randomized algorithms $\mathcal{A}$.

We assume that one is happy to obtain an ε -approximation to the solution with a probability $0 < P < 1$. If we allow equality, i.e., $0 < P \leq 1$ in Definition 1, then one may use R_n as an accuracy measure for both randomized and deterministic algorithms. In such a way it is consistent to consider a wider class $\mathcal{A}$ of algorithms that contains both classes: randomized and deterministic algorithms.

Definition 2. Consider the set $\mathcal{A}$ of algorithms A :

$$\mathcal{A} = \{A : Pr(R_n \leq \varepsilon) \geq c\}$$

that solve a given problem with a probability error R_n such that the probability that R_n is less than a priori given constant ε is bigger than a constant $c < 1$.

In such a setting it is correct to compare randomized algorithms with algorithms based on low discrepancy sequences like Sobol's $\Lambda\Pi_\tau$ ¹ sequences.

3 The Algorithm

The algorithm we study is based on Sobol's $\Lambda\Pi_\tau$ sequences. A Sobol's vector of dimension s ($\Lambda\Pi_\tau$ point) is considered as a centrum of a sphere with a radius ρ . Then a random point uniformly distributed on the sphere is taken and a random variable is defined as a value of the integrand at that random point.

To describe the algorithm one should start with the definition of $\Lambda\Pi_\tau$ sequences. An *uniformly distributed sequence* (u.d.s.) of non-random points was introduced by Hermann Weyl in 1916 [14]. Denote by $S_n(\Omega)$ the number of points with $1 \leq i \leq n$ that lie in Ω , where $\Omega \subset \mathbf{E}^s$. The sequence $x^{(1)}, x^{(2)}, \dots$ is called an u.d.s. if, for an arbitrary region Ω ,

$$\lim_{n \rightarrow \infty} [S_n(\Omega)/n] = V(\Omega),$$

where $V(\Omega)$ is the s -dimensional volume of Ω .

Theorem 1. ([13,14])

The relation

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(\xi^{(j)}) = \int_{\mathbf{E}^s} f(x) dx \quad (3)$$

holds for all Riemann integrable functions f if and only if the sequence $x^{(1)}, x^{(2)}, \dots$ is u.d.s.

¹ We use Cyrillic letters to denote $\Lambda\Pi_\tau$ sequences instead of widely used now " LP_τ " notation to express our deep respect to Professor Ilya Meerovich Sobol' who used Cyrillic letters to define his sequences in his original work [9] written in Russian.

Comparing (1) with (3) one can conclude that if the random points $\xi^{(i)}$ are replaced by the points $x^{(i)}$ of u.d.s., then for a wide class of functions f the averages converge. In this case the i -th trial should be carried out using Cartesian coordinates $(x_{i,1}, \dots, x_{i,s})$ of the point $x^{(i)}$, rather than the random numbers $\gamma_{i,1}, \dots, \gamma_{i,s}$.

For practical purposes an u.d.s. must be found that satisfied three additional requirements [10,12]:

- (i) the best asymptote as $n \rightarrow \infty$;
- (ii) well distributed points for small n ;
- (iii) a computationally inexpensive algorithm.

All Π_τ -sequences given in [12] satisfy the first requirement. Good distributions like Π_τ sequences are also called (t, m, s) -nets and (t, s) -sequences in base b . To introduce them, define first an elementary s -interval in base b as a subset of $\mathbf{E}^s$ of the form

$$\prod_{j=0}^s \left[\frac{a_j}{b^{d_j}}, \frac{a_j + 1}{b^{d_j}} \right],$$

where a_j, d_j are integers and $a_j < d_j$ for all $j \in \{1, \dots, s\}$. Given 2 integers $0 \leq t \leq m$, a (t, m, s) -net in base b is a sequence $x^{(i)}$ of b^m points of $\mathbf{E}^s$ such that $\text{Card } P \cap \{x^{(1)}, \dots, x^{(b^m)}\} = b^t$ for any elementary interval P in base b of hypervolume $\lambda(P) = b^{t-m}$.

Given a non-negative integer t , a (t, s) -sequence in base b is an infinite sequence of points $x^{(i)}$ such that for all integers $k \geq 0, m \geq t$, the sequence $\{x^{(kb^m)}, \dots, x^{((k+1)b^m-1)}\}$ is a (t, m, s) -net in base b .

I.M. Sobol' defines his Π_τ -meshes and Π_τ sequences, which are (t, m, s) -nets and (t, s) -sequences in base 2 respectively. The terms (t, m, s) -nets and (t, s) -sequences in base b (also called Niederreiter sequences) were introduced in 1988 by H. Niederreiter [7]. The term Sobol's sequences was introduced in late English-speaking papers in comparison with Halton, Faure and other low-discrepancy sequences.

To generate the j -th component of the points in a Sobol's sequence, we need to choose a primitive polynomial of some degree s_j over the Galois field of two elements $\text{GF}(2)$

$$P_j = x^{s_j} + a_{1,j}x^{s_j-1} + a_{2,j}x^{s_j-2} + \dots + a_{s_j-1,j}x + 1,$$

where the coefficients $a_{1,j}, \dots, a_{s_j-1,j}$ are either 0 or 1.

A sequence of positive integers $\{m_{1,j}, m_{2,j}, \dots\}$ are defined by the recurrence relation

$$m_{k,j} = 2a_{1,j}m_{k-1,j} \oplus 2^2a_{2,j}m_{k-2,j} \oplus \dots \oplus 2^{s_j}m_{k-s_j,j} \oplus m_{k-s_j,j},$$

where $\oplus$ is the bit-by-bit *exclusive-or* operator. The initial values $m_{1,j}, \dots, m_{s_j,j}$ can be chosen freely provided that each $m_{k,j}, 1 \leq k \leq s_j$, is odd and less than 2^k .

The so-called direction numbers $\{v_{1,j}, v_{2,j}, \dots\}$ are defined by $v_{k,j} = \frac{m_{k,j}}{2^k}$. Then the j -th component of the i -th point in a Sobol's sequence, is given by

$$x_{i,j} = i_1 v_{1,j} \oplus i_2 v_{2,j} \oplus \dots,$$

where i_k is the k -th binary digit of $i = (\dots i_3 i_2 i_1)_2$. Here the notation $(\bullet)_2$ denotes the binary representation of numbers.

Subroutines to compute these points can be found in [1,11]. The work [6] contains more details. If $x^{(i)} \equiv (x_{i,1}, x_{i,2} \dots x_{i,s})$ is the i -th $\Lambda\Pi\tau$ point in $\mathbf{E}^s$, then the i -th random point $\xi^{(i)}(\rho)$ with a probability density function $p(x)$ may be generated in the following way: $\xi^{(i)}(\rho) = x^{(i)} + \rho\omega^{(i)}$, where $\omega^{(i)}$ is a unique uniformly distributed vector in $\mathbf{E}^s$ (obviously, $\omega^{(i)} = \{\cos \phi_i, \sin \phi_i\}$ in $\mathbf{E}^2$). The general idea is that we take a Sobol's $\Lambda\Pi\tau$ point (vector) $x^{(i)}$ of dimension s . Then $x^{(i)}$ is considered as a centrum of a sphere with a radius ρ . A random point $\xi \in \mathbf{E}^s$ uniformly distributed on the sphere is taken. Consider a random variable θ such that $\theta = f(\xi)$. One can prove the following theorem.

Theorem 2. *The mathematical expectation of the random variable $\theta = f(\xi)$ is equal to the value of the integral (2), that is*

$$E\theta = S(f) = \int_{E^s} f(x)dx.$$

Proof. Consider random points $\xi(\rho) \in \mathbf{E}^s$. Assume $\xi(\rho) = x + \rho\omega$, where ρ is relatively small $\rho \ll \left[\frac{a_j}{2^{d_j}}, \frac{a_j+1}{2^{d_j}}\right]$, such that $\xi^{(i)}(\rho)$ is still in the same elementary s -interval $\mathbf{E}_i^s = \prod_{j=0}^s \left[\frac{a_j}{2^{d_j}}, \frac{a_j+1}{2^{d_j}}\right]$, where the pattern $\Lambda\Pi\tau$ point $x^{(i)}$ is. We use a subscript i in $\mathbf{E}_i^s$ to indicate that the i -th $\Lambda\Pi\tau$ point $x^{(i)}$ is in it. So, we assume that if $x^{(i)} \in \mathbf{E}_i^s$, then $\xi^{(i)}(\rho) \in \mathbf{E}_i^s$ too. Taking into account the latter fact probability density function $p(x)$ of the random variable $\xi^{(i)}(\rho)$ is constant. According to the definition $p(x)$ must be a non-negative function, such $\int_{E^s} p(x)dx = 1$. Thus, $p(x) = 1$ for $x \in E^s$. Now, for the mathematical expectation of $\theta = f(\xi)$ we have

$$E\theta = \int_{E^s} f(x)p(x)dx = \int_{E^s} f(x)dx.$$

The latter equality proves the theorem.

Theorem 2 allows to define a randomized algorithm. One can take the Sobol's $\Lambda\Pi\tau$ point $x^{(i)}$ and *shake* them a little bit. *Shaking* means to define random points $\xi^{(i)}(\rho) = x^{(i)} + \rho\omega^{(i)}$ according to the procedure described above. At first stage such a procedure seems similar to randomized quasi-Monte Carlo (see, for example [4,5]). But it is not. In randomized quasi-Monte Carlo people generally use random variable instead of fixed parameters in generators of quasi-random points. Here we use completely random points on spheres that have centurms in quasi-random points. We do not have an ambition to investigate different properties of the proposed algorithm and possible randomized quasi-Monte Carlo algorithm at this stage.

4 Numerical Tests

A number of numerical experiments have been performed to study numerically properties of the proposed algorithm. Our expectations based on theoretical results and a large number of numerical experiments are that for non-smooth functions our algorithm has advantages against quasi-Monte Carlo even for relatively low dimensions. Here we present some tests run for the following non-smooth integrand:

$$f_1(x_1, x_2, x_3, x_4) = \sum_{i=1}^4 |(x_i - 0.5)^{-1/3}|, \quad (4)$$

for which even the first derivative does not exist. Applications like that appear in some important problems in financial mathematics. The referent value of the integral $S(f_1)$ is approximately equal to 7.55953.

To make a comparison we also consider an integral with a smooth integrand:

$$f_2(x_1, x_2, x_3, x_4) = e^{x_1+2x_2} \frac{\cos(x_3)}{1 + x_2 + x_3 + x_4}. \quad (5)$$

The second integrand (5) is an infinitely smooth function with a referent value of the integral $S(f_2)$ approximately equal to 1.83690. The integration domain in both cases is $E^4 = [0, 1]^4$. In this work the algorithm with Gray code implementation and sets of direction numbers proposed by Joe and Kuo [3] for generating Sobol's quasirandom sequences are used. Our Monte Carlo algorithm (MCA) involves generating random points uniformly distributed on a sphere with a radius ρ . One of the best available random number generators, SIMD-oriented Fast Mersenne Twister (SFMT) [8] 128-bit pseudorandom number generator of period $2^{19937} - 1$ has been used to generate the required random points. The radius ρ depends on the integration domain, number of samples and minimal distance between Sobol's deterministic points δ . On the other hand, we observed experimentally that the behaviour of the relative error of numerical integration is significantly influenced by the fixed radius of spheres. That is why the values of the radius ρ are presented according to the number of samples n used in our experiments, as well as to a fixed coefficient, *radius coefficient* $\kappa = \rho/\delta$. The latter parameter gives the ratio of the radius to the minimal distance between Sobol's points (see, Table 1).

Table 1. Radius ρ of spheres of the random points

n	Min. dist., δ	κ	ρ	κ	ρ	κ	ρ
10	0.43301	0.001	0.00043	0.09	0.03897	0.4	0.17321
10^2	0.13166	0.001	0.00013	0.09	0.01185	0.4	0.05266
10^3	0.06392	0.001	0.00006	0.09	0.00575	0.4	0.02557
10^4	0.02812	0.001	0.00003	0.09	0.00253	0.4	0.01125
$5 \cdot 10^4$	0.01400	0.001	0.00001	0.09	0.00126	0.4	0.00560

Table 2. Relative error and computational time for numerical integration ($S(f_1) \approx 7.55953$)

n	SFMT		Sobol's alg.		MCA				
	Rel. err.	Time (s)	Rel. err.	Time (s)	δ	κ	ρ $\times 10^3$	Rel. err.	Time (s)
10^2	0.0114	0.01	0.0565	< 0.01	0.132	0.03	3.9	0.0038	0.01
						0.45	59	0.0050	0.01
10^3	0.0023	0.06	0.0114	0.01	0.064	0.03	1.9	0.0016	0.10
						0.45	29	0.0004	0.11
10^4	0.0006	0.53	0.0023	0.06	0.028	0.03	0.8	4e-05	3.56
						0.45	12.7	0.0002	3.58
$3 \cdot 10^4$	0.0002	1.63	0.0011	0.19	0.019	0.03	0.6	0.0002	28.5
						0.45	8.3	0.0003	28.8
$5 \cdot 10^4$	0.0009	2.67	0.0008	0.29	0.014	0.03	0.4	0.0002	74.8
						0.45	6.3	2e-05	75.7

The relative error (in absolute value) and the computational time for different values of the radius coefficient κ ($\kappa = 0.03$ and $\kappa = 0.45$) for numerical integration of the non-smooth integrand f_1 and for the smooth integrand f_2 are presented in Table 2 and Table 3 respectively. Relative error is the absolute error divided by the referent value. In the both tables the results with points obtained using SFMT generator have been denoted by “SFMT“, the results with points of Sobol's quasirandom sequences - by “Sobol's alg.“, and the results corresponding to the proposed Monte Carlo algorithm have been denoted by “MCA“. The proposed algorithm follows the relative error tendency of the original algorithm as one can expect. The numerical tests show certain advantages of the proposed Monte Carlo algorithm according to the estimated error in comparison to the original one, as well as to the algorithm using pseudorandom numbers generated by SFMT generator.

The computational time given in Table 2 and Table 3 has been estimated for obtaining the desired value after averaging on 10 algorithm runs. The computational complexity of the proposed algorithm does not increase essentially when the number of points n is not too large in comparison with the case when Sobol's sequences are used. For large values of n the complexity increases mainly because of the algorithm for finding the minimal distance between $\Lambda\Pi_\tau$ points in $\mathbf{E}^s$. This algorithm requires $O(n^2)$ operations. New random points for our algorithm have been generated using original Sobol's sequences and modeling a random direction in s -dimensional space. An additional computational time is needed to generate random points inside the integration domain. In case of unlikely event of generating random point outside the domain, the random point is rejected and a new random direction is chosen while the new random point gets into the domain.

Table 3. Relative error and computational time for numerical integration ($S(f_2) \approx 1.83690$)

n	SFMT		Sobol's alg.		MCA				
	Rel. err.	Time (s)	Rel. err.	Time (s)	δ	κ	ρ $\times 10^3$	Rel. err.	Time (s)
10^2	0.0350	< 0.01	0.0155	< 0.01	0.132	0.03	3.9	0.0160	0.01
						0.45	59	0.0264	0.01
10^3	0.0045	0.01	0.0023	< 0.01	0.064	0.03	1.9	0.0025	0.06
						0.45	29	0.0058	0.06
10^4	0.0016	0.10	0.0002	0.02	0.028	0.03	0.8	0.0003	3.29
						0.45	12.7	0.0016	3.28
3.10^4	0.0006	0.28	0.0001	0.04	0.019	0.03	0.6	0.0002	28.5
						0.45	8.3	0.0011	28.4
5.10^4	0.0004	0.46	6e-05	0.07	0.014	0.03	0.4	0.0001	76.0
						0.45	6.3	0.0008	76.1

The relative errors for numerical integration that have been obtained applying the proposed MCA and Sobol's algorithm are presented on Figure 1. The first plot on the figure is given in logarithmic scale by y-axis for clarity. The graphical representation of the numerical results illustrates smaller relative error obtained using the proposed algorithm for the most cases - for various number of samples and values of the radius. When the radius coefficient increases, the radius tends to minimal distance between the corresponding Sobol's points. On the other hand, samples size increase leads to decrease of this minimal distance. The obtained results show that optimal values and ratios between these parameters exist to define the smallest relative error of the proposed algorithm in comparison with the original quasi-Monte Carlo algorithm using Sobol's $AI\tau$ sequences. Our algorithm gives an approximation with a better precision for more of the cases or a relative error closed to the corresponding error of the original algorithm.

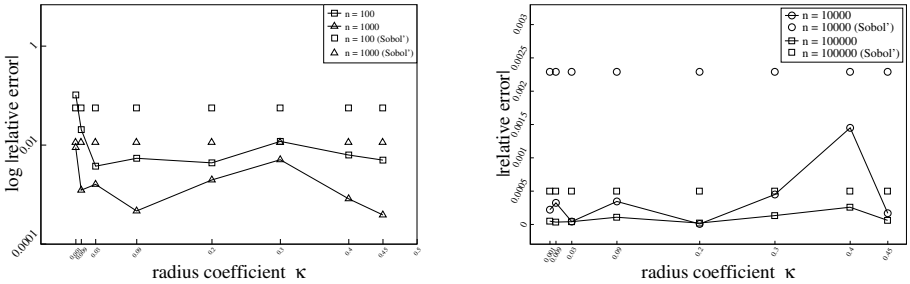


Fig. 1. Relative error according to the radius of the domain of random points

It is also confirmed that the difference of absolute values of the corresponding relatives errors tends to zero when sample size increases. Results confirming this fact are presented in Table 4.

Table 4. Difference of relative errors for Sobol's algorithm and the proposed MCA

$n \backslash \kappa$	0.009	0.03	0.2	0.45
10	0.07709	0.23746	0.20639	0.23037
10^2	0.03594	0.05277	0.05214	0.05155
10^3	0.01014	0.00976	0.00940	0.01099
10^4	0.00197	0.00225	0.00228	0.00212
$3 \cdot 10^4$	0.00102	0.00094	0.00084	0.00079
$5 \cdot 10^4$	0.00077	0.00062	0.00077	0.00078

5 Discussion of Applicability

One can clearly observe that the proposed algorithm improves the error estimates for non-smooth integrands when the radius ρ is smaller than the minimal distance between $\Lambda\Pi_\tau$ points δ . Strongly speaking the proposed approach is applicable if ρ is much smaller than δ . The implementation of the algorithm shows that this requirement is not very strong. Even for relatively large radiuses ρ the results are good. The reason is that centers of spheres are very well uniformly distributed by definition. So that, even for large values of radiuses of *shaking* the generated random points continue to be well distributed.

It should be mentioned here that for relatively low number of points (< 1000) the proposed algorithm gives results with a high accuracy. The relative error is approximately equal to 0.0038 for $n = 100$. For the same sample size the Sobol's algorithm gives more than 10 times higher error. For $n = 1000$ our algorithm gives relative error 0.0004 – 0.0016 depending on the parameter κ while the Sobol's algorithm gives 0.0114. This is an important fact because one has a possibility to estimate the value of the integral with a relatively high accuracy using a small number of random points.

If one deals with smooth functions, then the proposed algorithm is definitely better than the plain Monte Carlo based on SFMT generator, but it is not better than Sobol's algorithm based on $\Lambda\Pi_\tau$ points. Actually the results are very close to each other. This result is not unexpectable since the Sobol's algorithm is known to be very good for smooth functions (especially, for not very high dimensions).

6 Concluding Remarks

The proposed algorithm combines properties of two of the best available approaches - Sobol's quasi-Monte Carlo integration and a high quality pseudorandom number SFMT generator.

- The algorithm has advantages against quasi-Monte Carlo and SFMT for non-smooth integrands. For relatively small number of points the proposed approach gives much better results (smaller relative errors) than Sobol's quasi-Monte Carlo integration.
- In case of smooth functions the proposed algorithm has significant advantage against plain Monte Carlo that uses SFMT generator with respect to the relative error.

Acknowledgment

The research reported in this paper is partly supported by the Bulgarian NSF Grants DTK 02/44/2009, SuperCA++ (DCVP02/1) and DO 02-215/2008.

References

1. Bradley, P., Fox, B.: Algorithm 659: Implementing Sobol's Quasi Random Sequence Generator. *ACM Trans. Math. Software* 14(1), 88–100 (1988)
2. Dimov, I.T.: *Monte Carlo Methods for Applied Scientists*. World Scientific, London (2008)
3. Joe, S., Kuo, F.Y.: Constructing Sobol' Sequences with Better Two-dimensional Projections. *SIAM J. Sci. Comput.* 30, 2635–2654 (2008)
4. L'Ecuyer, P., Lecot, C., Tuffin, B.: A Randomized Quasi-Monte Carlo Simulation Method for Markov Chains. *Operations Research* 56(4), 958–975 (2008)
5. L'Ecuyer, P., Lemieux, C.: Recent Advances in Randomized Quasi-Monte Carlo Methods. In: Dror, M., L'Ecuyer, P., Szidarovszki, F. (eds.) *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pp. 419–474. Kluwer Academic Publishers, Boston (2002)
6. Levitan, Y., Markovich, N., Rozin, S., Sobol', I.: On Quasi-random Sequences for Numerical Computations. *USSR Comput. Math. and Math. Phys.* 28(5), 755–759 (1988)
7. Niederreiter, H.: Low-Discrepancy and Low-Dispersion Sequences. *Journal of Number Theory* 30, 51–70 (1988)
8. Saito, M., Matsumoto, M.: SIMD-oriented Fast Mersenne Twister: a 128-bit Pseudorandom Number Generator. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pp. 607–622. Springer, Heidelberg (2008)
9. Sobol', I.: *Multidimensional Quadrature Formulae and Haar functions (in Russian)*. Nauka, Moscow (1969)
10. Sobol', I.: *Monte Carlo Numerical Methods (in Russian)*. Nauka, Moscow (1973)
11. Sobol', I.: On the Systematic Search in a Hypercube. *SIAM J. Numerical Analysis* 16, 790–793 (1979)
12. Sobol', I.: On Quadratic Formulas for Functions of Several Variables Satisfying a General Lipschitz Condition. *USSR Comput. Math. and Math. Phys.* 29(6), 936–941 (1989)
13. Sobol', I.: Quasi - Monte Carlo Methods. In: Sendov, B.I., Dimov, I.T. (eds.) *International Youth Workshop on Monte Carlo Methods and Parallel Algorithms 1989*, pp. 75–81. World Scientific, Singapore (1990)
14. Weyl, H.: Ueber die Gleichverteilung von Zahlen mod Eins. *Math. Ann.* 77(3), 313–352 (1916)

Using Monte-Carlo Simulation for Risk Assessment: Application to Occupational Exposure during Remediation Works

M.L. Dinis and A. Fiúza

University of Porto, Faculty of Engineering, CIGAR
Rua Roberto Frias, 4200-465 Porto, Portugal

Abstract. The aim of this study was to apply the Monte-Carlo techniques to develop a probabilistic risk assessment. The risk resulting from the occupational exposure during the remediation activities of a uranium tailings disposal, in an abandoned uranium mining site, was assessed. A hypothetical exposure scenario was developed and two different pathways were compared: internal exposure through radon inhalation and external through gamma irradiation from the contaminated tailings material. The input variables, such as the inhalation rate and the external exposure parameters, were considered as specific probabilistic distributions, each one characterized by its central tendency and dispersion parameters. Using the cumulative distribution function, a probabilistic value for each variable can be generated using a single random number. Thus, this methodology allows performing a probabilistic risk assessment generating a risk distribution.

Keywords: Risk and dose assessment, uranium tailings disposal, Monte-Carlo simulation, occupational exposure.

1 Introduction

The uranium mining in Portugal took place at 62 different sites, mostly in small open pit exploitations although the larger ones were underground mines or a combination of both. Most of the mining sites are located in the districts of Guarda and Viseu (central-east Portugal). One of these sites was the Urgeiriça uranium mine which was considered to be the country's most important uranium exploitation. The Urgeiriça mine was active from 1913 to 2000 and the total uranium concentrate production reached about 4730 tons.

The uranium mining and processing operations in Portugal have left a legacy of considerable environmental contamination. The extensive treatment of uranium ores at the Urgeiriça processing plant led to the production of large amounts of solid wastes (tailings) that were deposited into open-air dams. The most voluminous tailing, with an estimated volume of $1\,390\,000 \pm 40\,000\, m^3$ and an area of 13.3 hectares, consists of the sludge produced in the milling facility [1].

This tailings pile, known as the "Old Dam", includes most of the radioisotopes of the uranium decay chains as well as other hazardous chemical elements resulting from the treatment process (acid leaching). Radium is of specific concern in the uranium tailings as it decays into radon, a radioactive gas which may cause lung cancer.

Since 1996 the Portuguese government had to deal with the decommissioning of the mines, mills and other facilities and the rehabilitation of the mining sites. In 2005, the Environmental Monitoring Programme became mandatory and legally enforced. The overall environmental remediation programme at the Urgeiriça mine was planned for completion before the end of 2007. However, the tailings pile "Old Dam" rehabilitation was only concluded in April 2008.

This work focuses on the potential occupational exposure during the "Old Dam" remediation works that occurred mainly in 2008. The "Old Dam" was one of the many sites with radioactive material to be rehabilitated. The others sites in the same region are scheduled to be intervened between 2010 and 2013.

The tailings are a source of external radiation and furthermore a powerful source of radon and dust that disperses in the atmosphere. During the remediation stage, radon inhalation may lead to significant occupational radiation exposure. Since site remediation was carried out at places with enhanced dose rates and high concentrations of airborne radioactivity (long-lived alpha particles, radon progenies), a large part of the workers were exposed to radiation mainly from three main exposure pathways: i) inhalation of radon decay products, ii) inhalation of dust-borne long-lived alpha emitters and iii) external radiation. The research described in this paper focuses, in particular, on radon inhalation and external (gamma) radiation.

2 Methods and Materials

2.1 Risk Assessment and Monte-Carlo Simulations

Risk assessment tools have been widely used to evaluate environmental contaminations and the effects on humans and ecosystems. Taken as an example radon exhalation from uranium tailings, the hazards of the exposure through the inhalation of radon come from its radioactive decay daughters. When radon is inhaled it decays into other radioactive products. These will dissipate their energy (alpha radiation) in the lung cells becoming a potential cause of lung cancer.

The probability of lung cancer occurrence depends on the amount of energy dissipated per unit mass. The amount of radon inhaled, and consequently, the amount of radon daughters inside the body, depends on the breathing rate. This factor will contribute to the estimative of the intake dose which is then combined with established factors (toxicity values or cancer slope factors) to determine the human health risk for that particular exposure.

The estimate of the intake dose requires data on the type and concentration of the contaminant together with many exposure input parameters. Generally some fixed values obtained from statistical analysis of the observed concentrations

for a given contaminant are combined with fixed standard values for exposure input parameters such as intake rates. The fixed input parameters are often chosen as the maximum value over a range of possible values to ensure that the estimative is on the "safe side". This is known as a based deterministic approach. The level of contamination (either measured or modeled) as well as exposure input parameters and toxicity values will always have inherent variability and uncertainty and these are not considered in a deterministic approach, which may result in an overestimate of the intake dose.

Probabilistic-based methods provide more realistic estimates using probability density functions for the input data instead of using fixed single values; for each parameter a probability density function is assigned. These distributions can take several forms (e.g. normal, lognormal, uniform, triangular, etc.).

A probabilistic methodology, such as Monte Carlo simulations, may be used to generate the cumulative intake dose, and then an intake dose value of the 90th to 99,9th percentiles of this distribution may be used for further risk assessment. The result will be a cumulative distribution of the intake dose that would, in a more realistic way, account for variability and uncertainty.

2.2 Occupational Exposure in Remediation/Rehabilitation Activities

The International Commission on Radiological Protection (ICRP) has established for occupational exposure an effective dose limit of 20 mSv per year, averaged over 5 years (100 mSv in 5 years), with the further provision that the effective dose should not exceed 50 mSv in any single year [2]. European legislation follows these limit values legislated in the Directive 96/29 EURATOM. This directive is designed to establish uniform safety standards to protect the health of workers and the general public against the dangers of ionizing radiation.

Although this directive came into force for European member states in 2000, Portugal was an exception. Portugal has notified transposition measures which were distributed in various legislative texts, instead of a coherent and consolidated legal framework. The European Commission considered that this made Portuguese legislation on radiation protection too complex and caused uncertainty for the citizens regarding the relevant transposition provisions [3]. In this way, Portugal was considered to have failed in fulfilling its obligations on basic safety standards for the health protection of workers and the general public from ionizing radiation.

After these events, EURATOM Directive was completely transposed to national law, in November 2008. Regarding these circumstances, in the previous context (before November 2008), the great majority of the rehabilitation works were done by workers that were officially non-radiologically exposed and consequently, their dose limit was the same as for the public, 1 mSv/year.

The U. S. Environmental Protection Agency (EPA) evaluates the risk due to radiation exposure as the carcinogenic slope factor, representing the lifetime excess total cancer risk per unit intake or exposure. The product of the cancer slope factor by the dose received estimates the risk for a member of the critical

Table 1. Gamma radiation emitters and dose coefficients

Radionuclide	Average soil concentration ($C_{soil,i}$) (Bq/kg)[7]	Dose coefficients ($DC_{ext,i}$) (Sv s^{-1})/(Bq m^{-2})[8]
^{235}U	483	1.48×10^{-16}
^{234}Th	6506	8.32×10^{-18}
^{226}Ra	3004	6.44×10^{-18}
^{210}Pb	3046	2.48×10^{-18}
^{137}Cs	9.90	2.85×10^{-19}
^{40}K	1738	1.46×10^{-16}

group due to their activity. This risk represents the probability of cancer inducing by this particular exposure, in excess relative to the background risk.

The acceptable risk is generally defined as 10^{-6} for the general public and 10^{-5} for occupational works. This means that an additional one case of cancer is accepted for population of 1 million or 100 000, being the general background risk around 20% for most of the industrialized countries. A risk level of 1 in a million, or 1 in one hundred thousand, also implies a likelihood that up to one person, out of one million (or 100 000) equally exposed people would contract cancer if they are exposed continuously (24 hours per day) to a specific radiation dose over 70 years (an assumed average lifetime).

The exposure scenario adopted in this study considers both internal and external exposure for estimating the exposure of the workers involved in the remediation activities of the tailings pile "Old Dam" and evaluates the health risks by means of a Monte-Carlo simulation. The estimative includes the dose and the associated risk from the activities. The dose assessment was done exclusively in a deterministic way while the risk assessment was done in a probabilistic based methodology. The critical receptor is represented by an average adult worker, involved in the remediation of the tailings, assuming an exposure during an 8-hour work day, 5 days per week, 48 weeks/year (accounting for the receptor being away on vacation for 4 weeks per year), during 3 years. It was also assumed that all the working time is spent outdoor [4]. The relevant pathways considered for the workers exposure are radon inhalation and gamma radiation from the tailings.

2.3 Sampling Methods

A radon survey over an area of 13.3 ha in the tailings pile and in its vicinity was carried out during two field campaigns in 2001. The first one was done in spring using 45 sampling points and the second one was done in summer using 22 sampling points. The sampling campaigns comprised various types of measurements, including radon exhalation rates (Bq $m^{-2} s^{-1}$) and radon concentration (Bq/ m^3).

The radon concentrations in the atmospheric air, measured at 1 m above the soil, ranged from 195 to 1205 Bq/ m^3 , with an average value of 557 Bq/ m^3 . In

the vicinity of the tailings pile the measured radon concentration varied from 50 to 930 Bq/m³ with an average value of 251 Bq/m³ [6].

To assess the external dose, measurements on the radionuclides gamma emitters were carried out [7]. The radionuclides gamma emitter concentrations in the soil were assessed by gamma-spectrometry and are presented in Table 1.

3 Applied Methodology

3.1 Effective Dose Assessment

The critical group for which individual doses are to be assessed is representative of the adult workers involved in the remediation activities. The effective dose due to radon inhalation may be calculated through the following equation:

$$D_{Rn} = C_{Rn} \times DC_{inh} \times E_f \times f_{eq}, \quad (1)$$

where D_{Rn} is the annual dose resulting from radon inhalation (mSv/year); C_{Rn} is the average radon concentration in breathing air at the tailings pile (Bq/m³); DC_{ing} is the radon effective dose equivalent factor (mSv/(Bq h m⁻³)); E_f is the outdoor exposure frequency (hour/year) and f_{eq} is the equilibrium factor for radon decay products (unitless).

We have adopted the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) recommendation for the conversion of potential alpha energy exposure (Bq h m⁻³) to effective dose equivalent (mSv) [5]. A value of 9×10^{-6} mSv per Bq h m⁻³ was adopted for the radon effective dose equivalent factor. This conversion factor has implicit an adult average breathing rate of 19.2 m³/d [5]. It was assumed an outdoor exposure frequency of 1920 hours per year and an equilibrium factor for radon decay products of 0.4 [5].

For the external exposure dose due to the contaminated ground surface ($D_{ext,i}$), the U.S. EPA dose coefficients (Table 1) were converted into the appropriated units by assuming a soil density of 1600 kg/m³ (ρ) and a soil depth contamination of 1 m (T_s) [8], in the following equation:

$$D_{ext,i} = C_{soil,i} \times DC_{ext,i} \times E_f \times 3600 \times \rho \times T_s, \quad (2)$$

where $C_{soil,i}$ is the radionuclide concentration in soil (Bq/kg), $DC_{ext,i}$ is the dose coefficient and the subscript “i” corresponds to each radionuclide (Table 1). In practice, doses obtained from the assessment of exposure from external radiation and from intake of radon are combined for the assessment of the value of total effective dose for demonstrating compliance with dose limits and constraints.

3.2 Risk Assessment

In a simplified approach, the annual risk incurred to a receptor by internal exposure due to radon inhalation may be estimated combining the radon concentration, the individual breathing rate, the exposure frequency and the radon cancer slope factor as given by equation (3):

$$R_{Rn} = C_{Rn} \times B_R \times RC_{inh} \times E_f \times f_{eq}, \quad (3)$$

where R_{Rn} is the annual risk resulting from radon inhalation; B_R is the breathing rate (log-normally distributed) at the exposure location (m^3/d) and RC_{inh} is the radon slope factor for inhalation (Risk/Bq); a value of 2.04×10^{-10} was adopted for this parameter [4].

A log-normal distribution of the daily breathing rate, normalized to the average body weight, was adopted based on values published for this log-normal distribution. We adopted the mean and the standard deviation designated by the ICRP [2]. For this distribution the mean, standard deviation, median and 95th percentile are respectively 16.45, 4.69, 16.32, 20.25 m^3/d .

The Monte Carlo methodology was used to generate an output cumulative distribution of the exposure risk: the Monte-Carlo, as a probabilistic method, performs the combination of probability distributions by numerical simulation and calculates the risk several thousand of times by generating random values for the input variables from the distribution function. This process was implemented using Matlab for programming an algorithm with different probability distribution for some of the variables involved in the risk calculations and performing about 30000 random generations.

In the algorithm the risk equation is expressed as a function of carcinogenicity (radon slope factor) and radon concentration, both as point values, and two exposure variables (breathing rate and exposure frequency) that are characterized by probability distributions functions (PDFs). The computer selects a value for each exposure variable from a specified PDF (log-normal for breathing rate and triangular for exposure frequency) and calculates the corresponding risk. This process is repeated many times (30000), each time saving the set of input values and corresponding estimate of risk. The results from each simulation are displayed in a graph, in the form of a probability density function or the corresponding cumulative distribution function.

The cancer risk induced by external radiation was estimated using the external radionuclide slope factor for each one of the radionuclides contributing to the external gamma radiation exposure. The following equation was used to assess the resulting risk,

$$R_{ext} = \sum_{i=1}^n \times (C_{soil,i} \times T_e \times SF_{ext,i} \times E_f \times S_f). \quad (4)$$

The input parameters are the soil concentration for each radionuclide, $C_{soil,i}$ (Bq/kg), the gamma exposure time factor, T_e (8h/24h), the external radionuclide slope factor, $SF_{ext,i}$ (Risk/year)/(Bq/kg) [9], the external exposure frequency, E_f (1920 h/365 d) and the outdoor gamma shielding factor, S_f (1) [9].

4 Results and Discussion

For the hypothetical exposure scenario, the effective dose for one year of radon internal exposure at 557 Bq/ m^3 is 3.85 mSv while for external exposure the estimated total gamma radiation emitters dose is 4.5 mSv/year. The value for total effective dose from internal and external exposure is 8.35 mSv/year.

Table 2. Risk assessment summary results

Risk	Average $\pm\sigma$	Median	95th percentile
Annual Rn inhalation	$6.00 \times 10^{-5} \pm 6.24 \times 10^{-11}$	5.95×10^{-5}	7.38×10^{-5}
Gamma radiation	4.33×10^{-5}	-	-
Total	$1.03 \times 10^{-4} \pm 6.24 \times 10^{-11}$	1.03×10^{-4}	1.17×10^{-4}
During rehabilitation	$3.52 \times 10^{-4} \pm 3.16 \times 10^{-32}$	3.52×10^{-4}	3.52×10^{-4}
Incremental lifetime risk	$0.0072 \pm 3.06 \times 10^{-7}$	0.0072	0.0082

A summary of the risk assessment values for the hypothetical scenario created is presented in Table 2. For analyzing the results, the mean and median values, as well as the 95th percentile, were extracted and presented. The estimative includes: i) the resulting annual risk; ii) the risk incurred by the exposure during the period of time necessary to complete the rehabilitation works which was considered to be 3 years and iii) the incremental lifetime cancer risk.

The total annual risk incurred by external exposure and by radon inhalation is log-normally distributed with a mean, standard deviation, median and 95th percentile as presented in Table 2, being σ the standard deviation.

For dose assessment, external exposure to gamma radiation was found to be the most significant exposure pathway, although radon inhalation contributes up to 46% to the annual effective dose. However, the higher risk in this exposure scenario is associated with radon inhalation contributing up to 52% to the total risk. As mentioned before, deterministic dose assessment may overestimate dose values as it uses single input values, which are often chosen as the maximum over the range of possible values, and consequently higher uncertainty. Risk assessment probabilistic approach generates values with lower uncertainty as it uses parameters distributions, instead of single input values. According to radon concentration measured in this site and dose measurements only for external exposure, radon inhalation was effectively the most concern for radiological exposure. In the present study only the breathing rate probability distribution was used in the risk calculation, while the other input parameters were considered as constants. It was our intention to demonstrate that the highest contribution to the dose, the external gamma irradiation, does not correspond to the highest probabilistic risk originated by radon inhalation and that an assessment based only on doses deterministic estimative may imply a non-realistic situation. Probabilistic risk calculations should also be taken into consideration when assessing human health exposure. Further work is being developed for using probabilistic distributions for all uncertain or variable input parameters.

5 Conclusions

The present study is based on a standard occupational exposure scenario for workers involved in the remediation activities in a low level radioactive waste disposal. A deterministic approach is used to perform a dose assessment. The results predict that external exposure is the highest concerning pathway. A single

point estimative for the dose is obtained and is useful to compare with the legal limit values. However, the probabilistic risk estimative showed that internal exposure poses, in fact, the highest risk to the exposed workers, and this fact should be considered. In this context, the workers involved in the remediation of the "Old Dam" could have been subject to radiation exposure (through internal and external pathways) and, for the exposure scenario created, the assessed values involve a hypothetical meaningful risk. The results for incremental cancer risk during the mean lifetime (0.0072) can be expressed as a probability of seven chances in 1000 for a specific worker experiencing a cancer fatality as a result of this particular exposure. This value is added to the background risk. In addition, dust inhalation containing radionuclides was not considered and this fact may contribute significantly to a higher inhalation dose and consequently increase the risk.

Further remediation works are planned to take place from 2010 to 2013 for other uranium tailings contaminated sites. It is clear that in the previous remediation activities workers were radiologically exposed. This preliminary study shows that many safety measures must be accomplished in future works, either by implementing individual protection equipment (internal and external exposure) or by periodical monitoring and control, in order to assess the experienced working exposure.

References

1. Dinis, M.L., Fiúza, A.: Integrated Methodology for the Environmental Risk Assessment of an Abandoned Uranium Mining Site. In: Uranium, Mining and Hydrogeology, pp. 163–176. Springer, Berlin (2008) ISBN: 978-3-540-87745-5
2. ICRP, International Commission on Radiological Protection: The 2007 Recommendations of the International Commission on Radiological Protection, Publication 103, Annals ICRP, vol. 37(2-4). Elsevier, Amsterdam (2007)
3. E.U., European Commission: Commission takes legal action against Portugal regarding safety standards for ionizing radiation, IP/07/1527 (2007)
4. EPA, U.S. Environmental Protection Agency: Health Effects Assessment Summary Tables, Radionuclide Carcinogenicity Slope Factors: HEAST (1995)
5. EPA, U.S. Environmental Protection Agency: Exposure Factors Handbook, Office of Research and Development, EPA/600/P-95/002 (1997)
6. EXMIN: Estudo Director de Áreas de Minérios Radioactivos - 2 fase. Companhia de Indústria e Serviços Mineiros e Ambientais, SA (2003)
7. Falcão, J.M., Carvalho, J.P., Boavida, M.G., Leite, M.M.: MinUrar, Minas de Urânio e seus Resíduos: Efeitos na Saúde da População, Relatório científico I, Publ. INSA, INETI, ITN (2005)
8. Eckerman, K.F., Ryman, J.C.: Federal guidance Report n.12, Exposure-to-Dose Coefficients for General Application, Based on the 1987 Federal Radiation Protection Guidance, EPA-402-R-93-081 (1993)
9. EPA, U.S. Environmental Protection Agency: Cancer Risk Coefficients for Environmental Exposure to Radionuclides, Federal Guidance Report N 13, EPA 402-R-99-001, Office of Radiation and Indoor Air (1999)

The b -adic Diaphony as a Tool to Study Pseudo-randomness of Nets

Ivan Lirkov¹ and Stanislava Stoilova²

¹ Institute of Information and Communication Technologies
Bulgarian Academy of Sciences

Acad. G. Bonchev, bl. 25A, 1113 Sofia, Bulgaria

`ivan@parallel.bas.bg`

² Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Acad. G. Bonchev, bl. 8, 1113 Sofia, Bulgaria

`stoilova@math.bas.bg`

Abstract. We consider the b -adic diaphony as a tool to measure the uniform distribution of sequences, as well as to investigate pseudo-random properties of sequences. The study of pseudo-random properties of uniformly distributed nets is extremely important for quasi-Monte Carlo integration. It is known that the error of the quasi-Monte Carlo integration depends on the distribution of the points of the net. On the other hand, the b -adic diaphony gives information about the points distribution of the net.

Several particular constructions of sequences (x_i) are considered. The b -adic diaphony of the two dimensional nets $\{y_i = (x_i, x_{i+1})\}$ is calculated numerically. The numerical results show that if the two dimensional net $\{y_i\}$ is uniformly distributed and the sequence (x_i) has good pseudo-random properties, then the value of the b -adic diaphony decreases with the increase of the number of the points. The analysis of the results shows a direct relation between pseudo-randomness of the points of the constructed sequences and nets and the b -adic diaphony as well as the discrepancy.

1 Introduction

The quasi-Monte Carlo methods can be described in simple words as a deterministic version of the Monte Carlo methods. The difference lies in the replacement of the random points by well-distributed deterministic nodes. One way to introduce the randomization into the quasi-Monte Carlo method is randomizing the deterministic integration nodes used in the method, e.g. [9,10,2]. In this way we can combine the faster convergence rates of the quasi-Monte Carlo methods and the possibility to estimate the error of the Monte Carlo methods. In this context, the study of pseudo-random properties of the deterministic sequences is inescapable. The approach to examine the pseudo-randomness of the deterministic sequences is to estimate the measures of distribution of the points of these sequences.

Many authors study pseudo-random properties of the sequences by using the bounds of the discrepancy [3,4,5,11,12,13,14,1]. We consider the relationship between the pseudo-randomness and the b -adic diaphony.

We will recall some known notions. Let $\xi = (\mathbf{x}_j)_{j \geq 1}$ be a sequence in $[0, 1]^s$. For arbitrary integer M and $J = [\alpha, \beta) \subseteq [0, 1]^s$ $A_M(\xi; J)$ is the number of belonging to J points of ξ . We denote the Lebesgue measure on J with $\mu(J) = \prod_{i=1}^s (\beta_i - \alpha_i)$.

Definition 1. *The sequence $\xi = (\mathbf{x}_j)_{j \geq 1}$ is uniformly distributed mod 1 in $[0, 1]^s$ if for every $J = [\alpha, \beta) \subseteq [0, 1]^s$ the equality*

$$\lim_{M \rightarrow \infty} \frac{A_M(\xi; J)}{M} = \mu(J)$$

is hold, see [15].

The last equality shows when a sequence of points in $[0, 1]^s$ is uniformly distributed but it does not allow to compare the distributions of two sequences. For that purpose measures of the distribution of the sequences are used. Such measures are the discrepancy [8] and the diaphony [6].

Definition 2. *Let $M \geq 1$ be an arbitrary fixed integer and $\xi_M = \{\mathbf{x}_0, \dots, \mathbf{x}_{M-1}\}$ is a net of real numbers in $[0, 1]^s$. The quantity*

$$D(\xi_M) = \sup_{J \subseteq [0, 1]^s} \left| \frac{A(\xi_M; J)}{M} - \mu(J) \right|$$

is called a discrepancy of the given net.

Let $b \geq 2$ be a fixed integer and b denotes the base of the number system everywhere in this paper. Also, let an arbitrary $x \in [0, 1]$ have a b -adic representation $x = \sum_{l=0}^{\infty} x_l b^{-1-l}$, where for $l \geq 0$, $x_l \in \{0, 1, \dots, b-1\}$ and for infinitely many values of l , $x_l \neq b-1$. The integer part of b -adic logarithm of x is $\lfloor \log_b x \rfloor = -g$, if $x_l = 0$ for $l < g$ and $x_g \neq 0$. We denote the operation $x \dot{-} y = \sum_{l=0}^{\infty} [(x_l - y_l) \pmod{b}] b^{-1-l}$ and for vectors $\mathbf{x}, \mathbf{y} \in [0, 1]^s$ we note $\mathbf{x} \dot{-} \mathbf{y} = (x_1 \dot{-} y_1, x_2 \dot{-} y_2, \dots, x_s \dot{-} y_s)$, where $\mathbf{x} = (x_1, x_2, \dots, x_s)$, $\mathbf{y} = (y_1, y_2, \dots, y_s)$.

We define the functions $\gamma : [0, 1] \rightarrow \mathbb{R}$ and $\Gamma : [0, 1]^s \rightarrow \mathbb{R}$ as

$$\gamma(x) = \begin{cases} b+1 - (b+1)b^{1+\lfloor \log_b x \rfloor}, & \text{if } x \in (0, 1) \\ b+1, & \text{if } x = 0 \end{cases}$$

and

$$\Gamma(\mathbf{x}) = -1 + \prod_{d=1}^s \gamma(x_d), \quad \mathbf{x} = (x_1, x_2, \dots, x_s).$$

Definition 3. *The b -adic diaphony $F_M(\xi)$ of the first M elements of the sequence $\xi = (\mathbf{x}_i)_{i \geq 0}$ in $[0, 1]^s$ is defined as*

$$F_M(\xi) = \left(\frac{1}{(b+1)^s - 1} \frac{1}{M^2} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \Gamma(\mathbf{x}_i \dot{-} \mathbf{x}_j) \right)^{\frac{1}{2}},$$

where the coordinates of all points of the sequence ξ are b -adic rational.

Let $i = \sum_{l=0}^{\infty} i_l b^l$ be the b -adic representation of the non-negative integer i . Then the i -th element of the Van der Corput sequence is defined as

$$\zeta_b(i) = \sum_{l=0}^{\infty} i_l b^{-l-1}.$$

Let us introduce $f_M \equiv a_2 y_i^2 + a_1 y_i + a_0 \pmod{M}$, where $M \geq 2$ is an integer and a_2, a_1, a_0 are three integer parameters. The number M is called modulus. Let $y_0 \in [0, M)$ be initial starting point. The sequence of pseudo-random numbers $\left\{x_i = \frac{y_i}{M}\right\}$ is produced by quadratic congruential generator

$$y_{i+1} = f_M(y_i), \quad y_i \in [0, M), \quad i = 0, 1, \dots, M-1. \quad (1)$$

The quadratic congruential generator (1) is introduced by Knuth [7]. He also proved that the sequence y_i is purely periodic with maximum possible period length M if and only if:

$$(a_0, M) = 1;$$

$$p|a_2 \text{ for every prime } p|M, \quad p > 2;$$

$$a_1 \equiv 1 \pmod{p} \text{ for every prime } p|M, \quad p > 2;$$

$$\text{If } 9|M, \text{ then either } 9|a_2 \text{ or } a_1 \equiv 1 \pmod{9} \text{ and } a_2 a_0 \equiv 6 \pmod{9};$$

$$\text{If } 4|M, \text{ then } 2|a_2 \text{ and } a_2 \equiv a_1 - 1 \pmod{4};$$

$$\text{If } 2|M, \text{ then } a_2 \equiv a_1 - 1 \pmod{2}.$$

Some authors researched pseudo-randomness of $x_i, i = 0, 1, \dots, M-1$ under the discrepancy D_M of the two-dimensional net

$$(x_i, x_{i+1}) = \left(\frac{y_i}{M}, \frac{y_{i+1}}{M} \right), \quad i = 0, 1, \dots, M-1.$$

J. Eichenauer-Herrmann and H. Niederreiter [4,5] proved bounds of the discrepancy D_M of the two-dimensional net produced by quadratic congruential generator which are $D_M = \mathcal{O}\left(\frac{(\log M)^2}{\sqrt{M}}\right)$. Using the geometric approach

O. Blažeková and O. Strauch [1] obtained order $\mathcal{O}\left(\frac{(\log M)^{3/2}}{\sqrt{M}}\right)$ of the star-discrepancy D_M^* of the same net. From the uniform distribution theory [8] it is well known that the discrepancy and the star discrepancy are always of the same order of magnitude, they differ at most by 2^s , where s is the dimension. Obviously, the order obtained in [1] is better than the previously proved estimates in [4,5].

2 The b -adic Diaphony and Pseudo-randomness

The study of the pseudo-random property of the sequence $x_i, i = 0, 1, \dots$ is associated with an estimation of the distribution of the two-dimensional net (x_i, x_{i+1}) . Until now, the discrepancy is used to estimate the distribution of

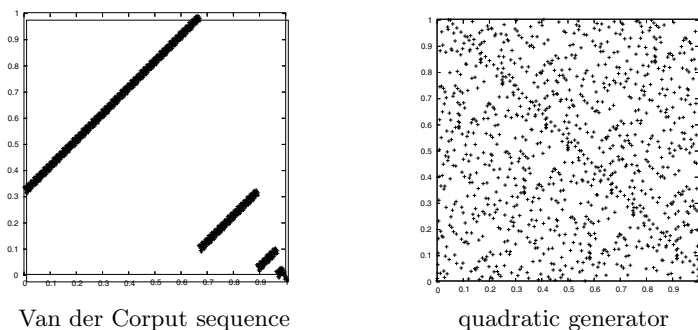


Fig. 1. The distribution of points of sequences (2) and (3), $M = 1024$, $b = 3$

the nets. Here we use the b -adic diaphony for the study of the distribution of the two-dimensional net (x_i, x_{i+1}) and the pseudo-randomness of the sequence x_i , $i = 0, 1, \dots$

2.1 Pseudo-randomness of the Van der Corput Sequence Using the b -adic Diaphony

We consider the net

$$(\zeta_b(i), \zeta_b(i+1)), \quad i = 0, 1, \dots, M-1. \quad (2)$$

This net is not uniformly distributed, because the points of the net lie on the lines $y = x + \frac{1}{b^{j+1}} + \frac{1}{b^j} - 1$, $j = 0, 1, 2, \dots$ (see Fig. 1).

The bad distribution of the two-dimensional net (2), based on the Van der Corput sequence is seen from the values of the b -adic diaphony in Table 1.

2.2 Pseudo-randomness of a Quadratic Generator Using the b -adic Diaphony

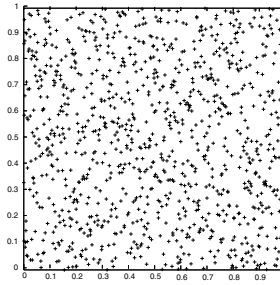
We consider the quadratic congruential generator (1) and obtain the sequence $x_i = \frac{y_i}{M}$ of quadratic congruential pseudo-random numbers. To investigate

Table 1. The diaphony F_M of the Van der Corput sequence, $b = 3$

$M = b^\nu, 3 \leq \nu \leq 10$		$M = 2^\mu, 4 \leq \mu \leq 16$			
M	F_M	M	F_M	M	F_M
27	0.374992	16	0.387033	4096	0.372105
81	0.37243	32	0.376644	8192	0.372104
243	0.372141	64	0.373283	16384	0.372104
729	0.372108	128	0.372489	32768	0.372104
2187	0.372105	256	0.372197	65536	0.372104
6561	0.372104	512	0.372126		
19683	0.372104	1024	0.372112		
59049	0.372104	2048	0.372106		

Table 2. The diaphony F_M of the quadratic generator $b = 3$

$3x^2 + x + 2(\bmod M)$ $M = b^\nu, 3 \leq \nu \leq 10$		$6x^2 + 3x + 1(\bmod M)$ $M = 2^\mu, 4 \leq \mu \leq 16$			
M	F_M	M	F_M	M	F_M
27	0.214727	16	0.187028	4096	0.0125823
81	0.10644	32	0.150217	8192	0.00912462
243	0.0592701	64	0.105382	16384	0.00630444
729	0.0348591	128	0.0760402	32768	0.00436687
2187	0.0165547	256	0.0544161	65536	0.00314525
6561	0.0119346	512	0.0362376		
19683	0.0072553	1024	0.0242248		
59049	0.00361669	2048	0.0171268		

**Fig. 2.** The distribution of points of the combination of quadratic generator with Van der Corput sequence with $M = 1024$, $b = 3$, $f_M(i) \equiv 6i^2 + 3i + 1(\bmod M)$

pseudo-random property of the sequence x_i , we calculate the b -adic diaphony of the net

$$(x_i, x_{i+1}), i = 0, 1, \dots, M - 1 \quad (3)$$

for two concrete quadratic generators in the case when $M = b^\nu$ and $M = 2^\mu$ and Table 2 shows the results.

2.3 Pseudo-random Property of the Combination of the Van der Corput Sequence with a Quadratic Generator

O. Strauch proposed to combine the Van der Corput sequence with a quadratic generator. In such way, the obtained net has a better pseudo-random property than original sequences.

To improve the distribution of the two-dimensional net we combine the Van der Corput sequence $\zeta_b(i)$ with the quadratic generator $y_{i+1} = f_M(y_i)$. In this way we obtain the net

$$(\zeta_b(y_i), \zeta_b(y_{i+1})), i = 0, 1, \dots, M - 1. \quad (4)$$

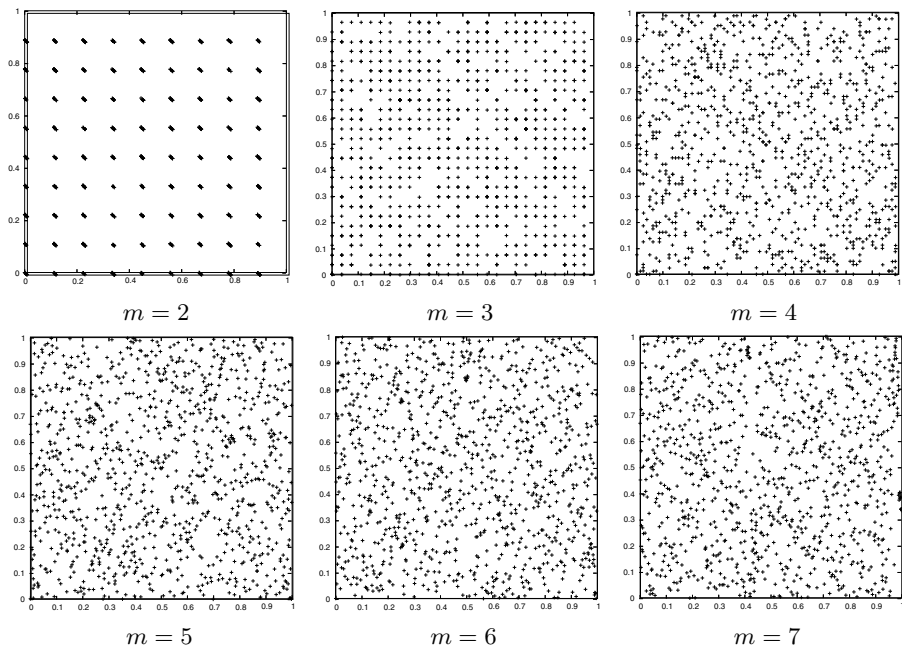


Fig. 3. The distribution of the combination of quadratic generator with Van der Corput sequence with $M = 1024$

Table 3. The diaphony F_M of the net (6) of the combination of quadratic generator $f_M(i) \equiv 6i^2 + 3i + 1 \pmod{M}$ with Van der Corput sequence, $b = 3$, $M = 2^\mu$, $4 \leq \mu \leq 16$, $3 \leq \nu \leq 9$, and $m \leq \nu$

M	F_M							
	m=2	m=3	m=4	m=5	m=6	m=7	m=8	m=9
16	0.19484	0.20886						
32	0.14907	0.1374	0.17695					
64	0.11024	0.09258	0.09042					
128	0.10049	0.06829	0.0731	0.07685				
256	0.08596	0.05764	0.04742	0.04808	0.05253			
512	0.07979	0.03917	0.03482	0.03402	0.03164			
1024	0.07538	0.03483	0.0256	0.02179	0.02098	0.03035		
2048	0.07313	0.0298	0.02051	0.01555	0.01698	0.01398		
4096	0.07217	0.02667	0.01602	0.01264	0.01263	0.01061	0.01293	
8192	0.07154	0.02542	0.01174	0.01009	0.00927	0.00878	0.00862	0.01095
16384	0.07118	0.02441	0.01042	0.00657	0.00639	0.00577	0.00602	0.00702
32768	0.07109	0.02385	0.00900	0.00517	0.00511	0.00458	0.00403	0.00416
65536	0.07102	0.02364	0.00841	0.00423	0.00347	0.00295	0.00305	0.00286

If the quadratic generator produced purely full period of the length M , then the net (4) has the same points as $(\zeta_b(i), \zeta_b(f_M(i))), i = 0, 1, \dots, M - 1$. The distribution of the obtained net is seen at Fig. 2.

Table 4. The diaphony F_M of the net (6) of the combination of quadratic generator $f_M(i) \equiv 3i^2 + i + 2 \pmod{M}$ with Van der Corput sequence, $b = 3$, $M = b^\nu$, $3 \leq \nu \leq 10$, and $m \leq \nu$

M	F_M							
	m=2	m=3	m=4	m=5	m=6	m=7	m=8	m=9
27	0.23612	0.37499						
81	0.20728	0.21735	0.37243					
243	0.11450	0.18947	0.21512	0.37214				
729	0.08438	0.08945	0.18739	0.21487	0.37211			
2187	0.07770	0.04994	0.08626	0.18715	0.21484	0.37211		
6561	0.07182	0.03851	0.04458	0.08590	0.18713	0.21484	0.37210	
19683	0.07173	0.02585	0.03139	0.04395	0.08585	0.18713	0.21484	0.37210
59049	0.07118	0.02562	0.01334	0.03050	0.04388	0.08585	0.18713	0.21483

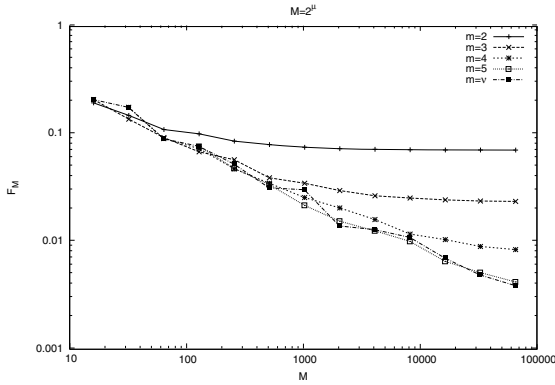


Fig. 4. The diaphony F_M of the combination of quadratic generator with Van der Corput sequence, $M = 2^\mu$

2.4 Simplification

For $x \in [0, 1)$ with the b -adic expression $x = 0.x_1x_2 \dots x_{m-1}x_mx_{m+1} \dots$ let $\zeta_{b^m}^*(x)$ be defined as $\zeta_{b^m}^*(x) = 0.x_mx_{m-1} \dots x_2x_1$. O. Strauch proposed the net

$$\zeta_{b^m}^* \left(\frac{y_i}{M} \right), i = 0, 1, \dots, M-1. \quad (5)$$

For pseudo-randomness of (5) we study the b -adic diaphony F_M of the two-dimensional net

$$\left(\zeta_{b^m}^* \left(\frac{y_i}{M} \right), \zeta_{b^m}^* \left(\frac{y_{i+1}}{M} \right) \right), i = 0, 1, \dots, M-1.$$

If $f_M(i)$ has a purely full period, then the net has the same points as

$$\left(\zeta_{b^m}^* \left(\frac{i}{M} \right), \zeta_{b^m}^* \left(\frac{f_M(i)}{M} \right) \right), i = 0, 1, \dots, M-1 \quad (6)$$

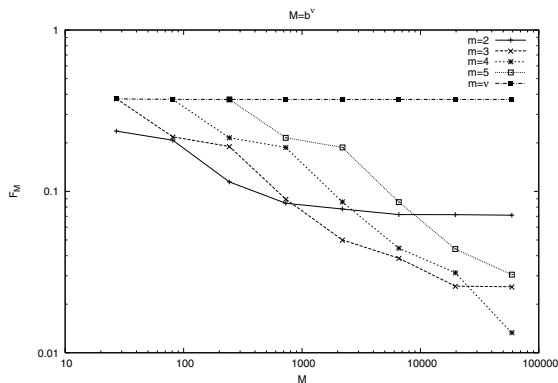


Fig. 5. The diaphony F_M of the combination of quadratic generator with Van der Corput sequence, $M = b^\nu$

and the same b -adic diaphony. The distribution of the points of the net (6) for six values of the number m is shown in Fig. 3.

Tables 3 and 4 as well as Fig. 4 and 5 show the computed b -adic diaphony of the nets using two quadratic generators with functions $f_M(i) \equiv 6i^2 + 3i + 1 \pmod{M}$, $M = 2^\mu$ and $f_M(i) \equiv 3i^2 + i + 2 \pmod{M}$, $M = 3^\nu$.

Conclusion and Future Work

The obtained results show that the b -adic diaphony is a good tool to study pseudo-randomness of sequences and nets. The calculations for the b -adic diaphony of the net (2) confirm the fact that the Van der Corput sequence is a deterministic and does not have pseudo-random properties. Last figures illustrate that the b -adic diaphony of the net (6) decreases with the increasing of the number of the points. This shows that the net (6) is uniformly distributed and therefore the sequence (5) has good pseudo-randomness. Hence, the b -adic diaphony can be used to research the pseudo-randomness of the sequences and nets. Furthermore, the b -adic diaphony of the nets (4) and (6) as well as of the sequence (5) can be theoretically estimated. In the future we plan to find such theoretical bounds.

Acknowledgments. We would like to thank Professor Oto Strauch for the wonderful ideas about the combination of the Van der Corput sequence with quadratic generator and the simplification of this combination. The study of pseudo-randomness of the proposed by Prof. Oto Strauch sequences is very interesting and useful for us. The authors thank to Professor Ivan Dimov for very useful remarks during the work on the paper. This work is supported by the project Bg-Sk-207, Bulgarian NSF.

References

1. Blažeková, O., Strauch, O.: Pseudo-randomness of quadratic generators. *Uniform Distribution Theory* 2(2), 105–120 (2007)
2. Dimov, I., Atanassov, E.: Exact Error Estimates and Optimal Randomized Algorithms for Integration. In: Boyanov, T., Dimova, S., Georgiev, K., Nikolov, G. (eds.) *NMA 2006*. LNCS, vol. 4310, pp. 131–139. Springer, Heidelberg (2007)
3. Drmota, M., Tichy, R.F.: *Sequences, Discrepancies and Applications*. LNM, vol. 1651. Springer, Heidelberg (1997)
4. Eichenauer-Herrmann, J., Niederreiter, H.: On the discrepancy of quadratic congruential pseudorandom numbers. *J. Comput. Appl. Math.* 34(2), 243–249 (1991)
5. Eichenauer-Herrmann, J., Niederreiter, H.: An improved upper bound for the discrepancy of quadratic congruential pseudorandom numbers. *Acta Arithmetica* 69(2), 193–198 (1995)
6. Grozdanov, V., Stoilova, S.: The b -adic diaphony. *Rendiconti di Matematica* 22, 203–221 (2002)
7. Knuth, D.E.: *Seminumerical algorithms*, 2nd edn. The art of computer programming, vol. 2. Addison Wesley, Reading (1981)
8. Kuipers, L., Niederreiter, H.: *Uniform distribution of sequences*. John Wiley, New York (1974)
9. L'Ecuyer, P., Lemieux, C.: Recent Advances in Randomized Quasi-Monte Carlo Methods. In: Dror, M., L'Ecuyer, P., Szidarovszki, F. (eds.) *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pp. 419–474. Kluwer Academic Publishers, Dordrecht (2002)
10. Lemieux, C., L'Ecuyer, P.: Randomized Polynomial Lattice Rules for Multivariate Integration and Simulation. *SIAM Journal on Scientific Computing* 24(5), 1768–1789 (2003)
11. Niederreiter, H.: Random number generation and quasi-Monte Carlo methods. In: *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 63. SIAM, Philadelphia (1992)
12. Niederreiter, H., Shparlinski, I.E.: On the distribution of inversive congruential pseudorandom numbers in parts of the period. *Mathematics of Computation* 70(236), 1569–1574 (2000)
13. Niederreiter, H., Shparlinski, I.E.: Exponential sums and the distribution of inversive congruential pseudorandom numbers with prime-power modulus. *Acta Arithmetica* XCII(1), 89–98 (2000)
14. Strauch, O., Porubský, Š.: *Distribution of Sequences: A Sampler*, Peter Lang, Frankfurt am Main (2005)
15. Weil, H.: Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* 77, 313–352 (1916)

Scatter Estimation for PET Reconstruction

Milan Magdics, Laszlo Szirmay-Kalos, Balazs Tóth,
Ádam Csenedesi¹, and Anton Penzov²

¹ Budapest University of Technology and Economics, Hungary

² Institute of Information and Communication Technologies, BAS, Bulgaria

Abstract. This paper presents a Monte Carlo scatter estimation algorithm for Positron Emission Tomography (PET) where positron-electron annihilations induce photon pairs that fly independently in the medium and eventually get absorbed in the detector grid. The path of the photon pair will be a *polyline* defined by the detector hits and scattering points where one of the photons changed its direction. The values measured by detector pairs will then be the total contribution, i.e. the integral of such polyline paths of arbitrary length. This integral is evaluated with Monte Carlo quadrature, using a sampling strategy that is appropriate for the graphics processing unit (GPU) that executes the process. We consider the contribution of photon paths to each pair of detectors as an integral over the Cartesian product set of the volume. This integration domain is sampled globally, i.e. a single polyline will represent all annihilation events occurred in any of its points. Furthermore, line segments containing scattering points will be reused for all detector pairs, which allows us to significantly reduce the number of samples. The scatter estimation is incorporated into a PET reconstruction algorithm where the scattered term is subtracted from the measurements.

1 Introduction

In positron emission tomography (PET) we need to find the spatial intensity distribution of positron–electron annihilations. During an annihilation event, two oppositely directed 511 keV photons are produced [Gea07]. We collect the number of simultaneous photon hits in detector pairs, also called *Lines Of Responses* or *LORs*: $(y_1, y_2, \dots, y_{N_{LOR}})$. The required output of the reconstruction method is the *emission density* function $x(\mathbf{v})$ that describes the number of photon pairs (i.e. the annihilation events) born in a unit volume around point $\mathbf{v}$. Tomography reconstruction algorithms are usually iterative. They start with an initial emission density, compute the detector response by simulating the photon transport and update the emission density taking into account the actual simulated and the measured detector responses [SV82]. Before being detected in the detectors, photons might interact with the matter in many ways, but in our energy range and for living organs only Compton scattering and the photoelectric absorption are relevant. The probability of scattering in unit distance is the *scattering cross section* σ_s . When scattering happens, there is a unique correspondence between the relative scattered energy and the cosine of the scattering angle θ , as defined

by the *Compton formula*:

$$\epsilon = \frac{1}{1 + \epsilon_0(1 - \cos \theta)},$$

where $\epsilon = E_1/E_0$ expresses the ratio of the scattered energy E_1 and the incident energy E_0 , and $\epsilon_0 = E_0/(m_e c^2)$ is the incident photon energy relative to the energy of the electron. The differential of the *scattering cross section*, i.e. the probability density that the photon is scattered from direction $\boldsymbol{\omega}'$ into differential solid angle $d\omega$ in direction $\boldsymbol{\omega}$, is given by the *Klein-Nishina formula* [Yan08]:

$$\frac{d\sigma_s(\mathbf{v}, \cos \theta, \epsilon_0)}{d\omega} = \frac{r_e^2 C(\mathbf{v})}{2} (\epsilon + \epsilon^3 - \epsilon^2 \sin^2 \theta),$$

where $\cos \theta = \boldsymbol{\omega} \cdot \boldsymbol{\omega}'$, $C(\mathbf{v})$ is the electron density, and $r_e = 2.82 \cdot 10^{-15}$ [m] is the classical electron radius. The Klein-Nishina formula defines the product of the scattering cross section $\sigma_s(\mathbf{v}, \epsilon_0)$ and the conditional probability density of the scattering direction. The scattering cross section can be obtained as the directional integral of the Klein-Nishina formula over the whole directional sphere:

$$\sigma_s(\mathbf{v}, \epsilon_0) = \int_{\Omega} \frac{d\sigma_s(\mathbf{v}, \cos \theta, \epsilon_0)}{d\omega} d\omega = \frac{r_e^2 C(\mathbf{v})}{2} \sigma_s^0(\epsilon_0) \quad (1)$$

where Ω is the directional sphere and σ_s^0 is the *normalized scattering cross section*:

$$\sigma_s^0(\epsilon_0) = \int_{\Omega} \epsilon + \epsilon^3 - \epsilon^2 \sin^2 \theta d\omega = -2\pi \int_{-1}^1 \epsilon + \epsilon^3 - \epsilon^2 \sin^2 \theta d \cos \theta.$$

The ratio of the Klein-Nishina formula and the scattering cross section is the *phase function*, which defines the probability density of the reflection direction, provided that reflection happens:

$$P_{KN}(\cos \theta, \epsilon_0) = \frac{d\sigma_s}{d\omega} / \sigma_s = \frac{\epsilon + \epsilon^3 - \epsilon^2 \sin^2 \theta}{\sigma_s^0(\epsilon_0)}.$$

The absorption cross section $\sigma_a(\epsilon_0)$ due to the photoelectric effect is approximately inversely proportional to the cube of the photon energy, thus

$$\sigma_a(\mathbf{v}, \epsilon_0) \approx \frac{\text{const}}{E^3} = \frac{\sigma_a(\mathbf{v}, 1)}{\epsilon_0^3}. \quad (2)$$

The proportionality ratio $\sigma_a(\mathbf{v}, 1)$ depends on the material compounds and grows rapidly (with a power between 4 and 5) with the atomic number of the elements.

2 Previous Work

A physically plausible scatter correction needs photon transport simulation and the evaluation of high-dimensional integrals in photon path space. As classical quadrature rules fail in higher dimensions due to the *curse of dimensionality*, these high-dimensional integrals are estimated by Monte Carlo or quasi-Monte Carlo methods [SK08]. Unfortunately, available Monte Carlo tools, like Geant4/GATE [Gea07, ABB⁺04], MCNP¹, SimSet², PeneloPET [EHV⁺06], are too general, and therefore not optimized for the particular task and not suitable for GPU execution. Thus, they are too slow to be incorporated into an on-line iterative reconstruction.

For effective simulation, we run our algorithm on the graphics processing unit (GPU), which is a massively parallel supercomputer. It can reach teraflops performance if its quasi-SIMD architecture is respected, i.e. if threads execute the same instruction sequence with no communication. The direct simulation of the photon transport would not meet this requirement since different photons may end up in the same detector which needs synchronized writes. Thus, we consider the adjoint problem and take a detector oriented viewpoint. For efficient evaluation, we transform the integral over the path space to a volumetric integral.

3 Scatter Estimation

If we consider photon scattering, the path of the photon pair will be a *polyline* containing the emission point somewhere inside one of its line segments (Fig. 1). This polyline includes scattering points $\mathbf{s}_1, \dots, \mathbf{s}_S$ where one of the photons changed its direction in addition to detector hit points $\mathbf{z}_1 = \mathbf{s}_0$ and $\mathbf{z}_2 = \mathbf{s}_{S+1}$. The values measured by detector pairs will then be the total contribution, i.e. the integral of such polyline paths of arbitrary length.

We consider the contribution of photon paths as an integral over the Cartesian product set of the volume. This integration domain is sampled globally, i.e. a single sample is used for the computation of all detector pairs. Sampling parts of photon paths globally and *reusing* a partial path for all detector pairs allow us to significantly reduce the number of samples.

To express the contribution of a polyline path, we take its line segments one-by-one and consider a line segment as a *virtual LOR* with two virtual detectors of locations, $\mathbf{s}_{i-1}$ and $\mathbf{s}_i$, and of differential areas projected perpendicularly to the line segment, $dA_{i-1}^\perp$ and $dA_i^\perp$ (Fig. 1). The contribution of a virtual LOR at its endpoints, i.e. the expected number of photon pairs going through $dA_{i-1}^\perp$ and $dA_i^\perp$ is $C(\mathbf{s}_{i-1}, \mathbf{s}_i)dA_{i-1}^\perp dA_i^\perp$, where contribution C is the product of several factors:

$$C(\mathbf{s}_{i-1}, \mathbf{s}_i) = G(\mathbf{s}_{i-1}, \mathbf{s}_i)X(\mathbf{s}_{i-1}, \mathbf{s}_i)T_1(\mathbf{s}_{i-1}, \mathbf{s}_i)B_1(\mathbf{s}_{i-1}, \mathbf{s}_i),$$

where $G(\mathbf{s}_{i-1}, \mathbf{s}_i)$ is the *geometry factor*, $X(\mathbf{s}_{i-1}, \mathbf{s}_i)$ is the *total emission* along the line segment, $T_{e_0}(\mathbf{s}_{i-1}, \mathbf{s}_i)$ is the *total attenuation due to out-scattering*, and

¹ <http://mcnp-green.lanl.gov/>

² http://depts.washington.edu/simset/html/simset_main.html

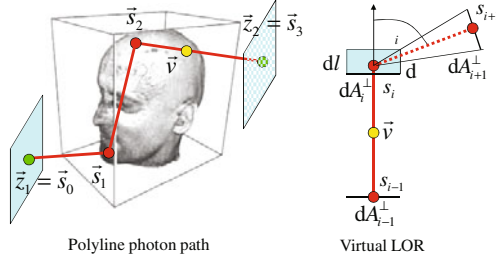


Fig. 1. The scattered photon path is a polyline (left) made of virtual LORs (right). The left figure depicts the case of $S = 2$.

$B_{\epsilon_0}(\mathbf{s}_{i-1}, \mathbf{s}_i)$ is the *total attenuation due to photoelectric absorption*, assuming photon energy ϵ_0 :

$$G(\mathbf{s}_{i-1}, \mathbf{s}_i) = \frac{1}{|\mathbf{s}_{i-1} - \mathbf{s}_i|^2}, \quad X(\mathbf{s}_{i-1}, \mathbf{s}_i) = \frac{1}{2\pi} \int_{\mathbf{s}_{i-1}}^{\mathbf{s}_i} x(\mathbf{l}) d\mathbf{l},$$

$$T_{\epsilon_0}(\mathbf{s}_{i-1}, \mathbf{s}_i) = e^{-\int_{\mathbf{s}_{i-1}}^{\mathbf{s}_i} \sigma_s(\mathbf{l}, \epsilon_0) d\mathbf{l}}, \quad B_{\epsilon_0}(\mathbf{s}_{i-1}, \mathbf{s}_i) = e^{-\int_{\mathbf{s}_{i-1}}^{\mathbf{s}_i} \sigma_a(\mathbf{l}, \epsilon_0) d\mathbf{l}}$$

In the line segment of the emission, the original photon energy has not changed yet, thus $\epsilon_0 = 1$.

Suppose that scattering happens around end point $\mathbf{s}_i$ of the virtual LOR in differential volume $d\mathbf{s}_i = dA_i^\perp d\mathbf{l}$, i.e. at run length $d\mathbf{l}$ (right of Fig. 1). Let us extend this virtual LOR by a single scattering step to form polyline $\mathbf{s}_{i-1}, \mathbf{s}_i, \mathbf{s}_{i+1}$. The probability that the photon scatters along distance $d\mathbf{l}$ and its new direction is in solid angle $d\omega$ is differential cross section $d\sigma_s(\mathbf{s}_i, \cos\theta_i, \epsilon_0^{(i)})/d\omega \cdot d\mathbf{l}$ where θ_i is the scattering angle. The scattered photon will go along virtual LOR $(\mathbf{s}_i, \mathbf{s}_{i+1})$ with differential area $dA_{i+1}^\perp$ at its end if area $dA_{i+1}^\perp$ subtends solid angle $d\omega$, that is:

$$d\omega = \frac{dA_{i+1}^\perp}{|\mathbf{s}_i - \mathbf{s}_{i+1}|^2}.$$

Upon scattering the photon changes its energy to

$$\epsilon_0^{(i+1)} = \frac{\epsilon_0^{(i)}}{1 + \epsilon_0^{(i)}(1 - \cos\theta)}.$$

This photon arrives at the other end of this virtual LOR if there is no further collision, which happens with probability $T_{\epsilon_0^{(i+1)}}(\mathbf{s}_i, \mathbf{s}_{i+1})B_{\epsilon_0^{(i+1)}}(\mathbf{s}_i, \mathbf{s}_{i+1})$.

Summarizing, the expected number of photon pairs born between $\mathbf{s}_{i-1}$ and $\mathbf{s}_i$ and reaching differential areas $dA_{i-1}^\perp$ and $dA_{i+1}^\perp$ via scattering at differential volume $d\mathbf{s}_i = d\mathbf{l} \cdot dA_i^\perp$ is:

$$C(\mathbf{s}_{i-1}, \mathbf{s}_i) \frac{d\sigma_s(\mathbf{s}_i, \cos\theta_i, \epsilon_0^{(i)})}{d\omega} T_{\epsilon_0^{(i+1)}}(\mathbf{s}_i, \mathbf{s}_{i+1}) B_{\epsilon_0^{(i+1)}}(\mathbf{s}_i, \mathbf{s}_{i+1}) dA_{i-1}^\perp d\mathbf{s}_i dA_{i+1}^\perp.$$

The integral of the contributions of paths of S scattering points is the product of these factors. For example, the integral of the contribution of paths of one scattering point is

$$\tilde{y}_L^{(1)} = \int_{D_1} \int_{D_2} \int_{\mathcal{V}} \cos \theta^{(0)} \cos \theta^{(2)} \frac{d\sigma_s(\mathbf{s}, \cos \theta, 1)}{d\omega} \mathcal{P}(\mathbf{z}_1, \mathbf{s}, \mathbf{z}_2) d\mathbf{s} dz_2 dz_1$$

where $\theta^{(0)}$ is the angle between the first detector's normal and the direction of $\mathbf{z}_1$ to $\mathbf{s}$, $\theta^{(2)}$ is the angle between the second detector's normal and the direction of $\mathbf{z}_2$ to $\mathbf{s}$, and $\mathcal{P}(\mathbf{z}_1, \mathbf{s}, \mathbf{z}_2)$ is the contribution of this polyline:

$$\mathcal{P}(\mathbf{z}_1, \mathbf{s}, \mathbf{z}_2) = C(\mathbf{z}_1, \mathbf{s}) T_{\epsilon_0}(\mathbf{s}, \mathbf{z}_2) B_{\epsilon_0}(\mathbf{s}, \mathbf{z}_2) + T_{\epsilon_0}(\mathbf{z}_1, \mathbf{s}) B_{\epsilon_0}(\mathbf{z}_1, \mathbf{s}) C(\mathbf{s}, \mathbf{z}_2). \quad (3)$$

The photon's energy level ϵ_0 is obtained from the Compton formula for scattering angle θ formed by directions $\mathbf{s} - \mathbf{z}_1$ and $\mathbf{z}_2 - \mathbf{s}$.

When the attenuation is computed, we should take into account that the photon energy changes along the polyline and the scattering cross section also depends on this energy, thus different cross section values should be integrated when the annihilations on a different line segment are considered. As we wish to reuse the line segments and not to repeat ray-marching redundantly, each line segment is marched only once assuming photon energy $\epsilon_0 = 1$, and attenuations T_1 and B_1 for this line segment is computed. Then, when the place of annihilation is taken into account and the real value of the photon energy ϵ_0 is obtained, initial attenuations T_1 and B_1 are transformed. The transformation is based on the decomposition of equations (1) and (2):

$$\sigma_s(\mathbf{l}, \epsilon_0) = \sigma_s(\mathbf{l}, 1) \cdot \frac{\sigma_s^0(\epsilon_0)}{\sigma_s^0(1)}, \quad \sigma_a(\mathbf{l}, \epsilon_0) = \frac{\sigma_a(\mathbf{l}, 1)}{\epsilon_0^3}.$$

Using this relation, we can write

$$\begin{aligned} T_{\epsilon_0} &= e^{-\int_{s_{i-1}}^{s_i} \sigma_s(\mathbf{l}, \epsilon_0) d\mathbf{l}} = e^{-\frac{\sigma_s^0(\epsilon_0)}{\sigma_s^0(1)} \int_{s_{i-1}}^{s_i} \sigma_s(\mathbf{l}, 1) d\mathbf{l}} = T_1^{\frac{\sigma_s^0(\epsilon_0)}{\sigma_s^0(1)}}. \\ B_{\epsilon_0} &= e^{-\int_{s_{i-1}}^{s_i} \sigma_a(\mathbf{l}, \epsilon_0) d\mathbf{l}} = e^{-\frac{1}{\epsilon_0^3} \int_{s_{i-1}}^{s_i} \sigma_a(\mathbf{l}, 1) d\mathbf{l}} = B_1^{\frac{1}{\epsilon_0^3}}. \end{aligned}$$

The energy dependence of the cross section $\sigma^0(\epsilon_0)$ is a scalar function, which can be pre-computed and stored in a table.

4 High-Dimensional Quadrature Computation

In the previous section we concluded that the scattered contribution is a sequence of increasing dimensional integrals. Numerical quadratures generate M discrete samples $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M$ in the domain of the integration and approximate the integral as:

$$\int f(\mathbf{u}) d\mathbf{u} \approx \frac{1}{M} \sum_{j=1}^M \frac{f(\mathbf{u}_j)}{p(\mathbf{u}_j)} \quad (4)$$

where $p(\mathbf{u}_j)$ is a density of samples. In the integral of the contribution, a sample $\mathbf{u}_j$ is a photon path connecting two detectors via S scattering points and containing an emission point somewhere:

$$\mathbf{u}_j = (\mathbf{s}_0^{(j)}, \mathbf{s}_1^{(j)}, \dots, \mathbf{s}_{S+1}^{(j)}) \quad \text{where } \mathbf{s}_0^{(j)} = \mathbf{z}_1 \text{ and } \mathbf{s}_{S+1}^{(j)} = \mathbf{z}_2.$$

For example, if $S = 1$ i.e. we consider single scattering, then $\mathbf{u}_j = (\mathbf{z}_1, \mathbf{s}^{(j)}, \mathbf{z}_2)$.

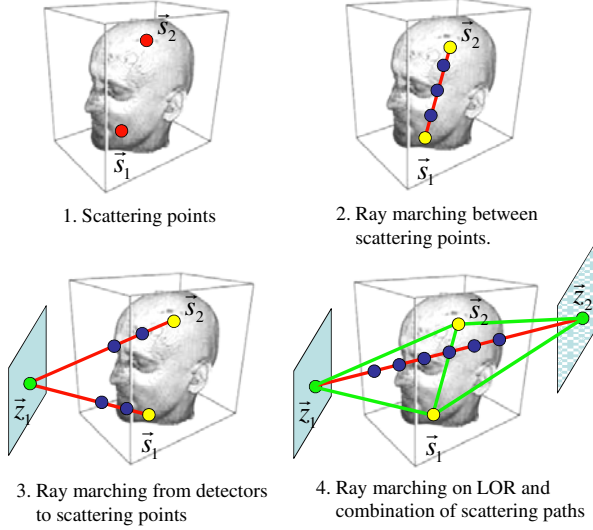


Fig. 2. Steps of the sampling process

As the computation of a single segment of such a path requires ray-marching and therefore is rather costly, we reuse the segments of a path in many other path samples. The basic steps of the path sampling process are shown by Fig. 2:

1. First, $N_{scatter}$ scattering points $\mathbf{s}_1, \dots, \mathbf{s}_{N_{scatter}}$ are sampled.
2. In the second step global paths are generated. If we decide to simulate paths of at most S scattering points, N_{path} ordered subsets of the scattering points are selected and paths of S points are established. If statistically independent random variables were used to sample the scattering points, then the first path may be formed by points $\mathbf{s}_1, \dots, \mathbf{s}_S$, the second by $\mathbf{s}_{S+1}, \dots, \mathbf{s}_{2S}$, etc. Each path contains $S - 1$ line segments, which are marched assuming that the photon energy has not changed from the original electron energy. Note that building a path of length S , we also obtain many shorter paths as well. A path of length S can be considered as two different paths of length $S - 1$ where one of the end points is removed. Taking another example, we get $S - 1$ number of paths of length 1. Concerning the cost, rays should be marched only once, so the second step altogether marches on $N_{path}(S - 1)$ rays.

3. In the third step, each detector is connected to each of the scattering points in a deterministic manner. Each detector is assigned to a computation thread, which marches along the connection rays. The total rays processed by the third step is $N_{det}N_{scatter}$.
4. Finally, detector pairs are given to GPU threads that compute the direct contribution and combine the scattering paths ending up in them. The direct contribution needs altogether $N_{detline}N_{LOR}$ ray-marching computations.

The described sampling process generates point samples. As these point samples are connected to all detectors, paths of length 2 (single scattering, $S = 1$) can be obtained from them. Paths longer than 2, i.e. simulating at least double scattering requires the formation of global paths. The integral quadrature of equation (4) is evaluated with these samples.

To reduce the variance of the random estimator, we should find a sampling density p that mimics the integrand. When inspecting the integrand, we should take into account that we evaluate a set of integrals (i.e. an integral for every LOR) using the same set of global samples, so the density should mimic the common factors of all these integrals. These common factors are the electron density $C(\mathbf{v})$ of the scattering points, so we mimic this function when sampling points. We store the scattering cross section at the energy level of the electron, $\sigma(\mathbf{v}, 1)$, which is proportional to the electron density. As the electron density function is provided by the CT reconstruction as a voxel grid, we, in fact, sample voxels. The probability density of sampling point $\mathbf{v}$ is:

$$p(\mathbf{v}) = \frac{\sigma_s(\mathbf{v}, 1)}{\int_V \sigma_s(\mathbf{v}, 1) d\mathbf{v}} = \frac{\sigma_s[V]}{C} \frac{N_{voxel}}{\mathcal{V}},$$

where $\sigma_s[V]$ is the scattering cross section at the energy level of the electron in voxel V , $C = \sum_{V=1}^{N_{voxel}} \sigma_s[V]$ is the sum of all voxels, and $\mathcal{V}$ is the volume of interest.

5 Results

The presented algorithm have been implemented in CUDA and run on nVidia GeForce 480 GFX GPUs. We have modeled the PET system of *NanoPET/CT* [Med] consisting of twelve square detector modules organized into a ring, and the system measures LORs connecting a detector to three other detectors being at the opposite sides of the ring, which means that $12 \times 3/2 = 18$ module pairs need to be processed. Each of the 12 detector modules consists of 81×39 crystals, thus $N_{det} = 12 \cdot (81 \times 39)$.

The computation effort can be analyzed by counting the number N_{ray} of rays needed to march on, which is

$$N_{ray} = N_{path}(S - 1) + N_{det}N_{scatter} + N_{detline}N_{LOR}.$$

In our particular case $S = 1$, $N_{scatter} = 128$, and $N_{detline} = 4$, thus — thanks to the heavy reuse of rays — scatter compensation requires just slightly more rays than the $N_{detline}N_{LOR}$ rays of the unscattered contribution computation.

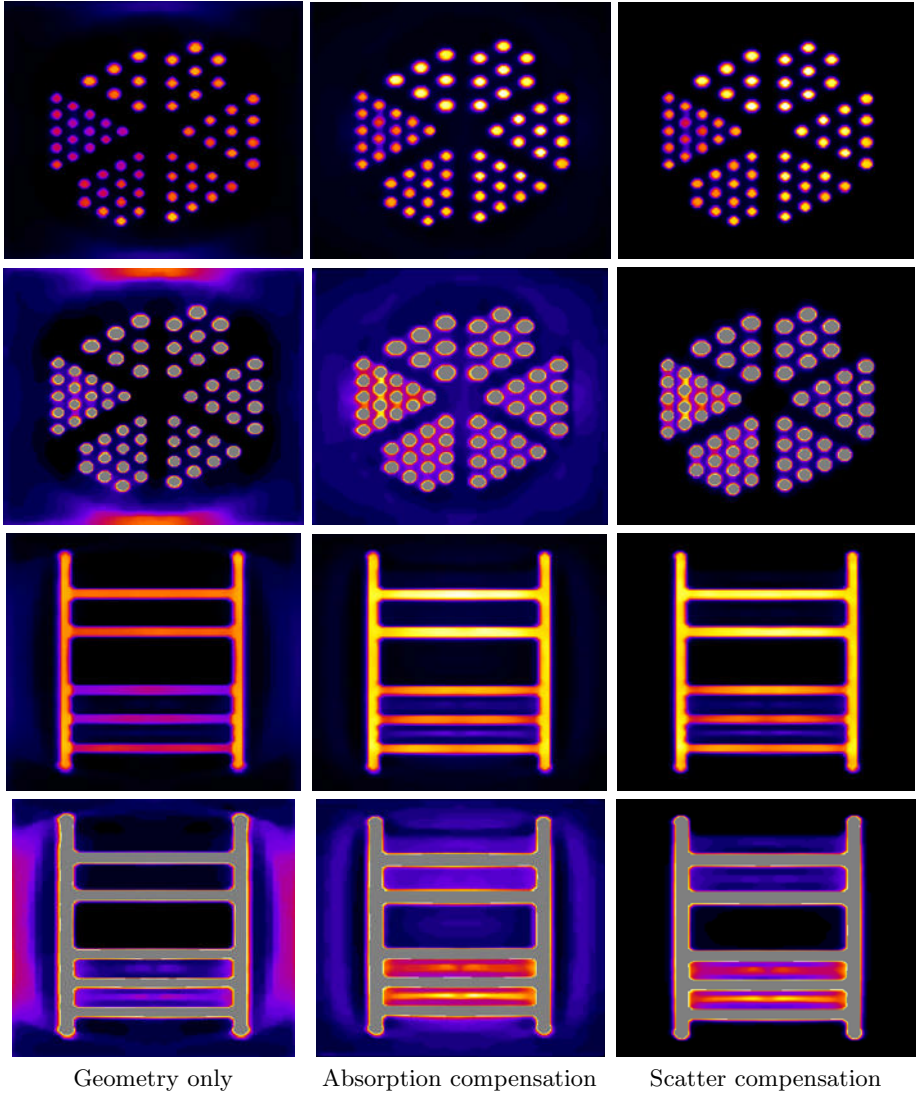


Fig. 3. Reconstruction results of the Derenzo phantom. The upper two rows depict a coronal and a sagittal slice of the reconstructed data, densities shown in the lower two rows are scaled by 5 in order to highlight the differences.

The reconstruction algorithm is an iteration of photon transfer simulation and density correction. We compared different options during the transfer simulation like computing only the geometry factors, adding the attenuation due to out-scattering and photoelectric absorption, and finally scattering compensation.

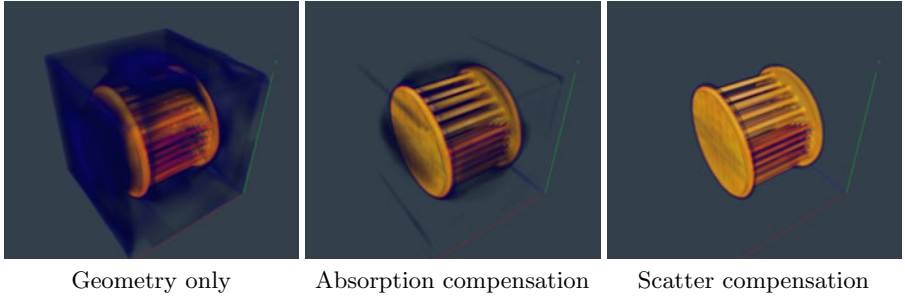


Fig. 4. 3D views of the Derenzo phantom reconstructions. We used a transfer function that emphasizes the cold noise in blue to make the differences more noticeable.

To compute single scattering, 128 scattering points are used, which are re-sampled in each iteration step. The algorithm has been tested on a Derenzo phantom that contains pipes with radioactive material. The Derenzo phantom is put in a cube of “super bone” of edge length 32 [mm]. Super bone has the same chemical compounds as the normal bone but it is ten times denser. In fact, it is even denser than steel, thus it can emphasize scattering and absorption phenomena. The results of the different options after 100 iteration steps are shown in Fig. 3 and Fig. 4. Note that getting the forward-projection to simulate more of the underlying physical process, the reconstruction can be made more accurate.

6 Conclusion

This paper proposed a GPU based scatter compensation algorithm for the reconstruction of PET measurements. The approach is restructured to exploit the massively parallel nature of GPUs. Based on the recognition that the requirements of the GPU prefer a detector oriented viewpoint, we solve the adjoint problem, i.e. originate photon paths in the detectors. The detector oriented viewpoint also allows us to reuse samples, that is, we compute many annihilation events with tracing a few line segments. The resulting approach can reduce the computation time of the fully 3D PET reconstruction to a few minutes.

Acknowledgement

This work has been supported by the TeraTomo project of the NKTH, OTKA K-719922 (Hungary), and Bulgarian NSF DTK 02/44. This work is connected to the scientific program of the “Development of quality-oriented and harmonized R+D+I strategy and functional model at BME” project. This project is supported by the New Hungary Development Plan (Project ID: TMOP-4.2.1/B-09/1/KMR-2010-0002).

References

- [ABB⁺04] Assié, K., Breton, V., Buvat, I., Comtat, C., Jan, S., Krieguer, M., Lazaro, D., Morel, C., Rey, M., Santin, G., Simon, L., Staelens, S., Strul, D., Vieira, J.-M., Walle, R.V.D.: Monte carlo simulation in PET and SPECT instrumentation using GATE. *Nuclear Instruments and Methods in Physics Research Section A* 527(1-2), 180–189 (2004)
- [EHV⁺06] Espana, S., Herraiz, J.L., Vicente, E., Vaquero, J.J., Desco, M., Udias, J.M.: PeneloPET, a Monte Carlo PET simulation toolkit based on PENELOPE: Features and validation. In: *IEEE Nuclear Science Symposium Conference*, pp. 2597–2601 (2006)
- [Gea07] Geant. Physics reference manual, Geant4 9.1. Technical report, CERN (2007)
- [Med] Mediso, <http://www.bioscan.com/molecular-imaging/nanopet-ct>
- [SK08] Szirmay-Kalos, L.: Monte-Carlo Methods in Global Illumination — Photo-realistic Rendering with Randomization. VDM, Verlag Dr. Müller, Saarbrücken (2008)
- [SV82] Shepp, L., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imaging* 1, 113–122 (1982)
- [Yan08] Yang, C.N.: The Klein-Nishina formula & quantum electrodynamics. *Lect. Notes Phys.*, vol. 746, pp. 393–397 (2008)

Modeling of the SET and RESET Process in Bipolar Resistive Oxide-Based Memory Using Monte Carlo Simulations

Alexander Makarov, Viktor Sverdlov, and Siegfried Selberherr

Institute for Microelectronics, TU Wien, Guhausstrae 27-29,
A-1040 Vienna, Austria
{makarov,sverdlov,selberherr}@iue.tuwien.ac.at

Abstract. A stochastic model of the resistive switching mechanism in bipolar oxide-based resistive random access memory (RRAM) is presented. The distribution of electron occupation probabilities obtained is in agreement with previous work. In particular, a low occupation region is formed near the cathode. Our simulations of the temperature dependence of the electron occupation probability near the anode and the cathode demonstrate a high robustness of the low occupation region. The RESET process in RRAM simulated with our stochastic model is in good agreement with experimental results.

Keywords: stochastic model, resistive switching, RRAM, Monte Carlo method.

1 Introduction

With memories based on charge storage (such as DRAM, flash memory, and other) approaching the physical limits of scalability, research on new memory structures has significantly accelerated. Several concepts as potential substitutes of the charge memory were invented and developed. Some of the technologies are already available as prototype (such as carbon nanotube RAM (NRAM), copper bridge RAM (CBRAM)), others as product (phase change RAM (PCRAM), magnetoresistive RAM (MRAM), ferroelectric RAM (FRAM), while the technologies of spin-torque transfer RAM (STTTRAM), racetrack memory, and resistive RAM (RRAM) are under research. A new type of memory must exhibit low operating voltages, low power consumption, high operation speed, long retention time, high endurance, simple structure, and small size [1].

One of the most promising candidates for future universal memory is the resistive random access memory (RRAM). It is based on new materials, such as metal oxides [2-4] and perovskite oxides [5]. This type of memory is characterized by high density, excellent scalability, low operating voltages (< 2 V), fast switching times (< 10 ns), and long retention time. On the other hand, RRAM devices have not demonstrated yet sufficient endurance. Unless this problem can be solved, this technology is unlikely to be brought to market in the 2020 time-frame [1]. Unfortunately, a proper fundamental understanding of the switching

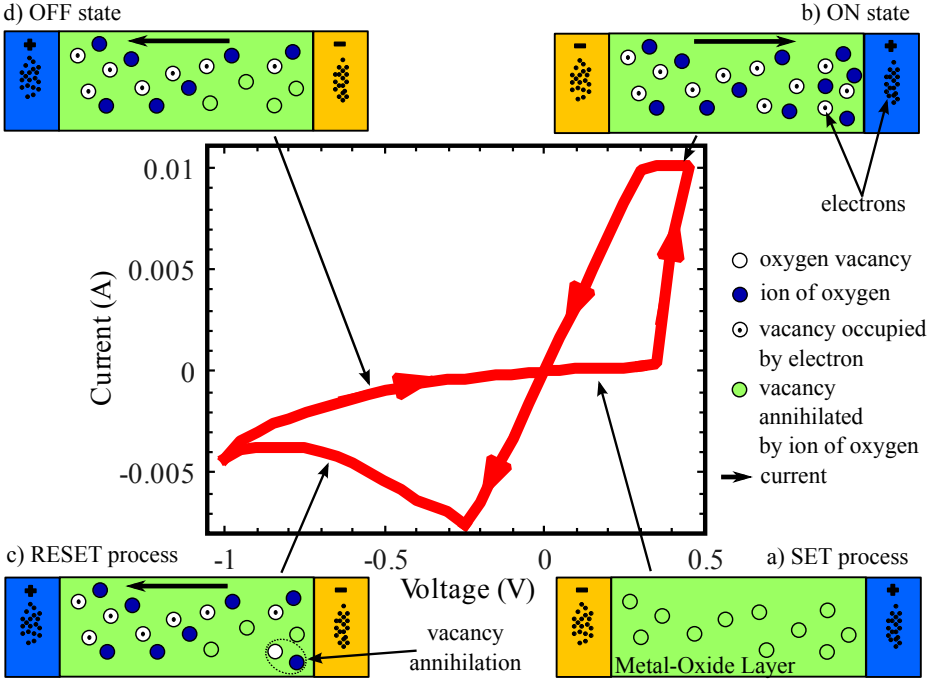


Fig. 1. Typical hysteresis cycle in RRAM and illustration of the resistive switching mechanism in bipolar oxide-based memory cell: (a) Schematic illustration of the SET process. (b) Schematic view of the conducting filament in the low resistance state (ON state). (c) Schematic illustration of the RESET process. (d) Schematic view of the conducting filament in the high resistance state (OFF state). Only the oxygen vacancies and ions that impact the resistive switching are shown.

mechanism in resistive random access memory (RRAM) is still missing, despite the fact that several physical mechanisms based on either electron or ion determined switching have been recently suggested in the literature: a model based on trapping of charge carriers [6], electrochemical migration of oxygen vacancies [7, 8], electrochemical migration of oxygen ions [9, 10], a unified physical model [11, 12], a domain model [13], a filament anodization model [14], a thermal dissolution model [15], and others.

In this work we present a stochastic model of the bipolar resistive switching mechanism based on electron hopping between the oxygen vacancies along the conductive filament in an oxide layer.

2 Model Description

We associate the resistive switching behavior in oxide-based memory with the formation and rupture of a conductive filament (CF) (Fig. 1).

The CF is formed by localized oxygen vacancies (V_o) [11, 12] or domains of V_o . Formation and rupture of a CF is due to a redox reaction in the oxide layer under a voltage bias. The conduction is due to electron hopping between these V_o .

For modeling the resistive switching in bipolar oxide-based memory by a Monte Carlo method, we describe the dynamics of oxygen ions (O^{2-}) and electrons in an oxide layer as follows:

- formation of V_o by O^{2-} moving to an interstitial position;
- annihilation of V_o by moving O^{2-} to V_o ;
- movement of O^{2-} between the interstitials;
- an electron hop into V_o from an electrode;
- an electron hop from V_o to an electrode;
- an electron hop between two V_o .

In order to model the dependences of transport on the applied voltage and temperature we choose the hopping rates for electrons as [16]:

$$\Gamma_{nm} = A_e \cdot \frac{dE}{1 - \exp(-dE/T)} \cdot \exp(-R_{nm}/a), \quad (1)$$

Here, A_e is a coefficient, $dE = E_n - E_m$ is the difference between the energies of an electron positioned at sites n and m , R_{nm} is the hopping distance, a is the localization radius. The hopping rates between an electrode (0 or $N + 1$) and an oxygen vacancy m are described as [12]:

$$\Gamma_m^{iC} = \alpha \cdot \Gamma_{0m}, \Gamma_m^{oC} = \alpha \cdot \Gamma_{m0}, \quad (2)$$

$$\Gamma_m^{iA} = \beta \cdot \Gamma_{(N+1)m}, \Gamma_m^{oA} = \beta \cdot \Gamma_{m(N+1)}, \quad (3)$$

Here, α and β are the coefficients of the boundary conditions on the cathode and anode, respectively, N is the number of sites, A and C stand for cathode and anode, and i and o for hopping on the site and out from the site, respectively.

To describe the motion of ions we have chosen the ion rates similar to (1):

$$\Gamma'_n = A_i \cdot \frac{dE}{1 - \exp(-dE/T)}, \quad (4)$$

Here we assume hopping only on a nearest interstitial. Thus, a distance-dependent term is included in A_i . dE includes the formation energy for the m -th V_o /annihilation energy of the m -th V_o , when O^{2-} is moving to an interstitial or back to V_o , respectively.

The current generated by hopping is calculated as:

$$I = q_e \cdot \sum dx / \sum \left(1 / \sum_m \Gamma_m \right) \quad (5)$$

Here q_e is the electron charge.

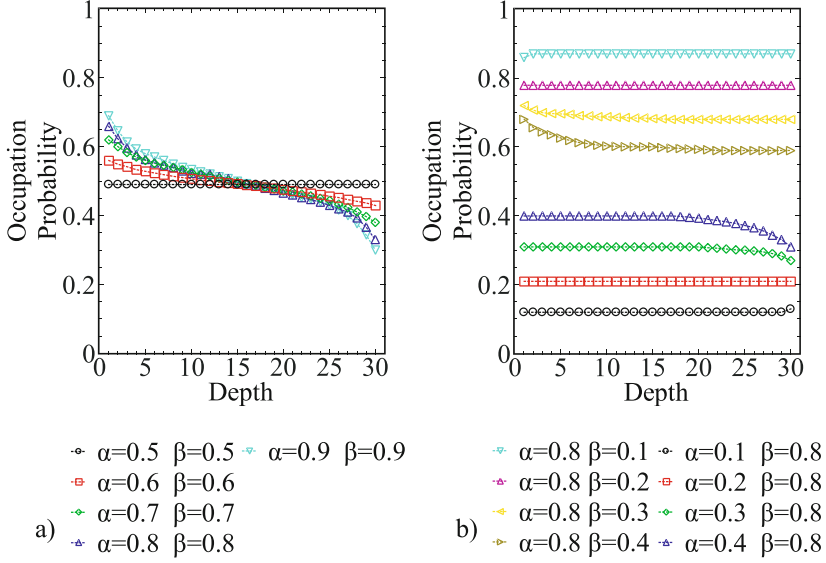


Fig. 2. Calculated distributions of electron occupation probabilities for unidirectional next nearest neighbor hopping between the V_o (the 1st V_o is near the cathode, the last V_o is near the anode): (a) $\alpha > 0.5$ and $\beta > 0.5$, $p_c = 0.5$; (b) $\beta < 0.5$ and $\beta < \alpha$, $p_c = 1 - \beta$; $\alpha < 0.5$ and $\alpha < \beta$, $p_c = \alpha$

3 Model Verification

Calculations are performed on one-dimensional lattices. All V_o are at the same energy level, if no voltage is applied. For simplify the calculations we assume that the oxygen vacancy is either empty or occupied by one electron.

3.1 Calculation of Electron Occupation Probabilities

To verify the proposed model, we first evaluate the average electron occupations of hopping sites under different conditions. For comparison with previous works all calculations in this subsection are made on a lattice consisting of thirty equivalent, equidistantly positioned hopping sites V_o .

Following [17], we first allow hopping in one direction and only to/from the closest V_o . The occupation probability of the central oxygen vacancies, p_c , is described depending on the boundary conditions as follows: 1) for $\alpha > 0.5$ and $\beta > 0.5$, $p_c = 0.5$; 2) for $\alpha < 0.5$ and $\alpha < \beta$, $p_c = \alpha$; 3) for $\beta < 0.5$ and $\beta < \alpha$, $p_c = 1 - \beta$. Fig.2 shows simulation results of our stochastic model, which are fully consistent with theoretical predictions [17].

To move from a model system [17] to a more realistic structure, we calculated the distribution of electron occupations for a chain, where hopping is allowed not only to/from the nearest V_o ($T = 0$, Fig. 3), and for systems, where hopping

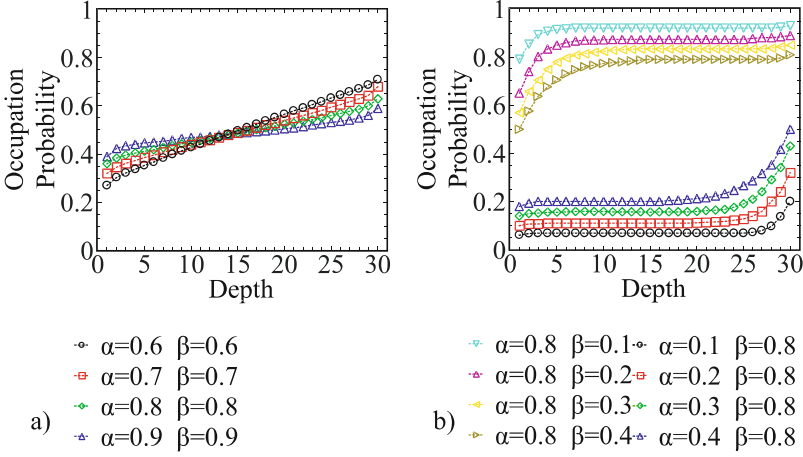


Fig. 3. Calculated distribution of electron occupation probabilities, if unidirectional hopping is allowed not only to/from the closest V_o ($T = 0$): (a) $\alpha > 0.5$ and $\beta > 0.5$; (b) $\beta < 0.5$ and $\beta < \alpha$; $\alpha < 0.5$ and $\alpha < \beta$

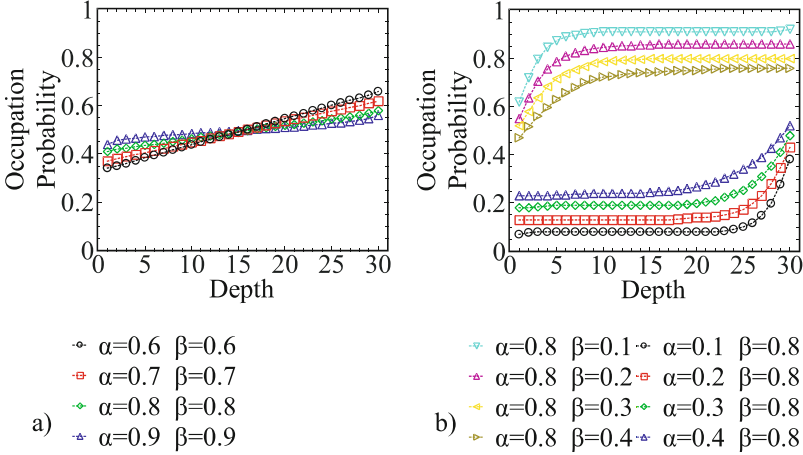


Fig. 4. Calculated distribution of electron occupation probabilities, for hopping according to (1-3), for $T > 0$: (a) $\alpha > 0.5$ and $\beta > 0.5$; (b) $\beta < 0.5$ and $\beta < \alpha$; $\alpha < 0.5$ and $\alpha < \beta$

(1-3) is allowed in both directions ($T > 0$, Fig. 4). Note that for $\alpha > 0.5$ and $\beta > 0.5$ (Fig. 3a and Fig. 4a) we still have $p_c = 0.5$ in the center, while for other values α, β we observe a decrease in p_c for $\alpha < \beta$ and an increase in p_c for $\beta < \alpha$.

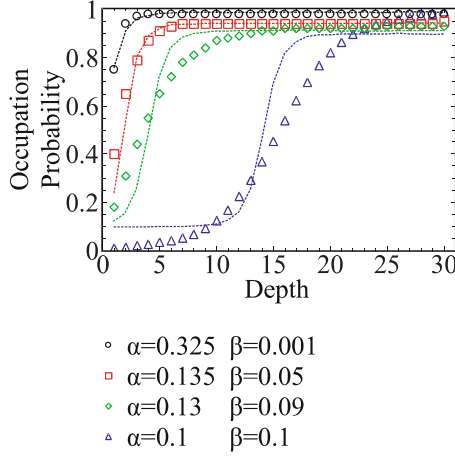


Fig. 5. Calculated distribution of electron occupation probabilities under different biasing voltages. Lines are from [12], symbols are obtained with our stochastic model.

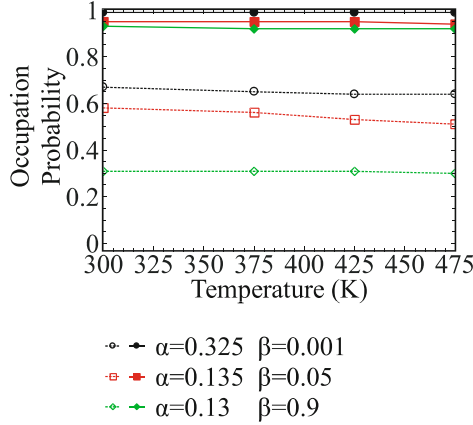


Fig. 6. Temperature dependence of electron occupation probability near the anode (line) and the cathode (dotted line)

We have calibrated our model in a manner to reproduce the results reported in [12], for $V = 0.6$ V to $V = 1.4$ V. Fig. 5 shows a case, when the hopping rate between two V_o is larger than the rate between the electrodes and V_o (i.e. $\alpha, \beta < 1$). In this case a low occupation region is formed near the cathode (bipolar behavior).

With the calibrated model we simulated the temperature dependence of the site occupations in the low occupation region. The results shown in Fig. 6 indicate high robustness of the low occupation region demonstrating changes of less than 10%, when the temperature is elevated from 25° C to 200° C.

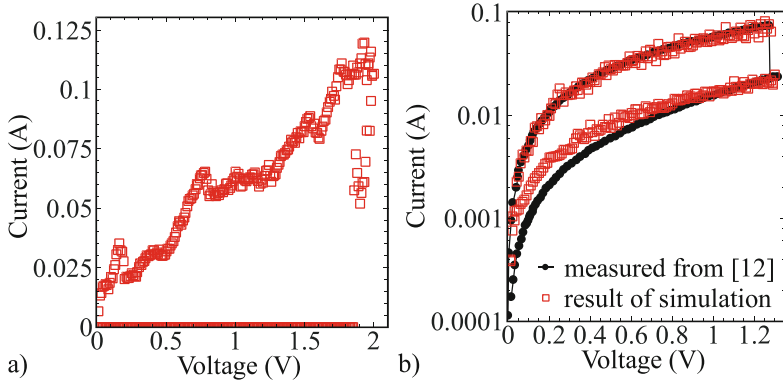


Fig. 7. $I - V$ characteristics for a single-CF device are obtained from our stochastic model: (a) SET $I - V$ characteristics; (b) RESET $I - V$ characteristics and measured results from [12]

3.2 Modeling of the SET and RESET Processes

For the simulations we have used a one-dimensional lattice consisting of thirty equivalent, equidistantly positioned hopping sites. To simplify calculations we assume that the coefficients of the boundary conditions are constant and equal to 0.1, independent of the applied voltage. In both simulations (SET and RESET process) we have used the same formation/annihilation energy for V_o . The result of the simulation of the SET process is shown in Fig. 7a.

To further demonstrate the capabilities of our model, we also simulated the RESET $I - V$ characteristics for a single-CF device [12]. For this purpose the CF was modified in such a way that for each V_o an oxygen ion is placed nearby. Fig.7b. shows the simulation result of the stochastic model, which is in perfect agreement with measurements from [12].

4 Conclusion

In this work we have presented a stochastic model of the bipolar resistive switching mechanism. The distribution of the electron occupation probabilities calculated with the model is in excellent agreement with previous work. The simulated RESET process in RRAM is in good agreement with the experimental result. The proposed stochastic model can be used for performance optimization of RRAM devices.

Acknowledgments. This research is supported by the European Research Council through the grant #247056 MOSILSPIN.

References

1. Kryder, M.H., Kim, C.S.: After Hard Drives - What Comes Next? IEEE Trans. on Mag. 45(10), 3406–3413 (2009)
2. Kugeler, C., Nauenheim, C., Meier, M., et al.: Fast Resistance Switching of TiO₂ and MSQ Thin Films for Non-Volatile Memory Applications (RRAM). In: NVM Tech. Symp., p. 6 (2008)
3. Chen, Y.S., Wu, T.Y., Tzeng, P.J.: Forming-free HfO₂ Bipolar RRAM Device with Improved Endurance and High Speed Operation. In: Symp. on VLSI Tech., pp. 37–38 (2009)
4. Dong, R., Lee, D.S., Xiang, W.F., et al.: Reproducible Hysteresis and Resistive Switching in Metal-Cu_xO-Metal Heterostructures. APL 90(4), 42107/1-3 (2007)
5. Lin, C.C., Lin, C.Y., Lin, M.H.: Voltage-Polarity-Independent and High-Speed Resistive Switching Properties of V-Doped SrZrO₃ Thin Films. IEEE Trans. on Electron Dev. 54(12), 3146–3151 (2007)
6. Fujii, T., Kawasaki, M., Sawa, A., et al.: Hysteretic Current-Voltage Characteristics and Resistance Switching at an Epitaxial Oxide Schottky Junction SrRuO₃/SrTi_{0.99}Nb_{0.01}O₃. APL 86(1), art. no. 012107 (2005)
7. Nian, Y.B., Strozier, J., Wu, N.J., et al.: Evidence for an Oxygen Diffusion Model for the Electric Pulse Induced Resistance Change Effect in Transition-Metal Oxides. PRL 98(14), 146403/1-4 (2007)
8. Wu, S.X., Xu, L.M., Xing, X.J.: Reverse-Bias-Induced Bipolar Resistance Switching in Pt/TiO₂/SrTi_{0.99}Nb_{0.01}O₃/Pt Devices. APL 93(4), 043502/1-3 (2008)
9. Szot, K., Speier, W., Bihlmayer, G., Waser, R.: Switching the Electrical Resistance of Individual Dislocations in Single-Crystalline SrTiO₃. Nature Materials 5, 312–320 (2006)
10. Nishi, Y., Jameson, J.R.: Recent Progress in Resistance Change Memory. In: Dev. Res. Conf., pp. 271–274 (2008)
11. Xu, N., Gao, B., Liu, L.F., et al.: A Unified Physical Model of Switching Behavior in Oxide-Based RRAM. In: Symp. on VLSI Tech., pp. 100–101 (2008)
12. Gao, B., Sun, B., Zhang, H., et al.: Unified Physical Model of Bipolar Oxide-Based Resistive Switching Memory. IEEE Electron Dev. Lett. 30(12), 1326–1328 (2009)
13. Rozenberg, M.J., Inoue, I.H., Sanchez, M.J.: Nonvolatile Memory with Multilevel Switching: A Basic Model. PRL 92(17), 178302-1 (2004)
14. Kinoshita, K., Tamura, T., Aso, H., et al.: New Model Proposed for Switching Mechanism of ReRAM. In: IEEE Non-Volatile Semicond. Memory Workshop 2006, pp. 84–85 (2006)
15. Russo, U., Ielmini, D., Cagli, C., et al.: Conductive-Filament Switching Analysis and Self-Accelerated Thermal Dissolution Model for Reset in NiO-Based RRAM. In: IEDM Tech. Dig., pp. 775–778 (2007)
16. Sverdlov, V., Korotkov, A.N., Likharev, K.K.: Shot-Noise Suppression at Two-Dimensional Hopping. PRB 63, 081302 (2001)
17. Derrida, B.: An Exactly Soluble Non-Equilibrium System: The Asymmetric Simple Exclusion Process. Phys. Rep. 301(1-3), 65–83 (1998)

Stochastic Algorithm for Solving the Wigner-Boltzmann Correction Equation

M. Nedjalkov¹, S. Selberherr¹, and I. Dimov²

¹ Institute for Microelectronics, TU Wien
Gußhausstraße 27-29/E360, A-1040 Vienna, Austria

² Institute for Parallel Processing, Bulgarian Academy of Sciences
Acad. G.Bontchev str Bl25A, 1113 Sofia, Bulgaria

Abstract. The quantum-kinetics of current carriers in modern nanoscale semiconductor devices is determined by the interplay between coherent phenomena and processes which destroy the quantum phase correlations. The carrier behavior has been recently described with a two-stage Wigner function model, where the phase-breaking effects are considered as a correction to the coherent counterpart. The correction function satisfies a Boltzmann-like equation.

A stochastic method for solving the equation for the correction function is developed in this work, under the condition for an a-priori knowledge of the coherent Wigner function. The steps of an almost optimal algorithm for a stepwise evaluation of the correction function are presented. The algorithm conforms the well established Monte Carlo device simulation methods, and thus allows an easy implementation.

1 Introduction

Modeling and simulation of electronic transport in semiconductor devices is challenged by the nanometer and picosecond scale processes which determine the functionality of modern integrated circuits. Quantum transport models are explored to correctly describe coherent processes, such as tunneling, in conjunction with de-coherence processes of scattering, which try to recover the classical behavior of the current carriers.

The Wigner-Boltzmann (WB) equation gives a comprehensive quantum-kinetic description of these phenomena, and has been recently applied for simulation of a variety of nanometer devices and involved transport phenomena [1]. Stochastic approaches to the WB equation efficiently describe the scattering processes, however, the coherent part of the transport is obtained at significant numerical costs. A scheme which uses coherent data obtained by alternative approaches has been developed recently. The scattering-induced correction to the coherent Wigner function satisfies a Fredholm integral equation of the second kind, with a free term determined by the coherent data.

Particle methods have been developed and used to calculate the free term. We have successfully applied these methods for very small devices, where this term can be regarded as a zeroth order correction. Here we utilize the numerical

Monte Carlo theory to derive a stochastic algorithm for solving the equation for the WB correction.

An important peculiarity is that the problem is comprised by two models with different dimensions: while the coherent transport involves two variables - the position and wave vector x, k_x , the scattering occurs in the three dimensional wave vector space, thus involving the transversal components $k_y, k_z = \mathbf{k}_\perp$. The two models are combined into a four dimensional space formulation by merely physical considerations. In this respect the sequel does not stick to the formal Monte Carlo schemes for solving integral equations, and in particular the adjoint equation, which proved as an already established approach to carrier transport problems [2], [3]. The adjoint equation remains rather implicit in the derivations, which refers to core schemes for solving integrals in favor of an emphasis on the physical aspects.

2 The Model

The time-independent Wigner-Boltzmann equation:

$$\frac{\hbar k_x}{m} \frac{\partial}{\partial x} f_w(x, k_x, \mathbf{k}_\perp) = \int dk'_x V_w(x, k'_x - k_x) f_w(x, k'_x, \mathbf{k}_\perp) + \int d\mathbf{k}' f_w(x, \mathbf{k}') S(\mathbf{k}', \mathbf{k}) - f_w(x, \mathbf{k}) \lambda(\mathbf{k}) \quad (1)$$

describes the coherent part of the carrier transport at a rigorous quantum level, accomplished by the Boltzmann scattering model of the phase-breaking processes. Here V_w is the Wigner potential, the Boltzmann scattering operator $S(\mathbf{k}, \mathbf{k}')$ presents the scattering rate for a transition from $\mathbf{k}$ to $\mathbf{k}'$. $\lambda(\mathbf{k}) = \int d\mathbf{k}' S(\mathbf{k}, \mathbf{k}')$ is the total out-scattering rate, so that the quantity S/λ is the probability density for scattering from the initial to the final state. The solution of (1) in the region D of a given device determines the physical characteristics of the current carriers and thus the circuit behavior of the device. The external factors which determine the solution are the applied bias, which controls the electric potential profile in the device, and the boundary conditions. The latter are assumed to satisfy the equilibrium distribution function deep inside the device leads. It is the Maxwell-Boltzmann distribution f_{MB} , which is the only function turning the second row in (1) to zero independently of the physical origin of the scattering processes.

The coherent problem is obtained from (1) by switching off all scattering processes. In this case the solution $f_w^c(x, k_x)$ does not depend on the transversal wave vector components. A proper alignment of the variables with the genuine problem must be such that f_w^c is recovered after an integration over the transversal ones. A consistent with the boundary condition assumption is the appearance of the equilibrium with respect to the transversal variables function $f_{MB}(\mathbf{k}_\perp)$:

$$f_w^c(x, \mathbf{k}) = f_w^c(x, k_x) \frac{\hbar^2}{2\pi m k T} e^{-\frac{\hbar^2 \mathbf{k}_\perp^2}{2m k T}} \quad (2)$$

This allows to define the function

$$f_w^\Delta(x, \mathbf{k}) = f_w(x, \mathbf{k}) - f_w^c(x, \mathbf{k}), \quad (3)$$

which is the scattering induced correction to the coherent Wigner function. The equation for the correction f_w^Δ is obtained by subtracting the coherent counterpart from (1). An immediate property of (3) is that the correction is zero at the device boundaries, where the same boundary conditions are assumed for both cases.

The Wigner potential is approximated by its classical limit valid for slowly varying potentials at a next step:

$$\int dk_x' V_w(x, k_x' - k_x) f_w^\Delta(x, k_x', \mathbf{k}_\perp) = -\frac{eE(x)}{\hbar} \frac{\partial f_w^\Delta(x, k_x, \mathbf{k}_\perp)}{\partial k_x} \quad (4)$$

This means that the force $F(x) = eE(x)$, given by the derivative of the potential, can be only a linear function within the spatial support of f_w^Δ , related to the spatial width of the electrons. Such an assumption in the general equation (1) precludes the quantum-mechanical description of the transport. The latter, however, has a different physical meaning in the equation for the correction. The width of the electron has been already accounted by the coherent solution, so that the limit precludes only correlations between the electric potential and the scattering processes.

The obtained model for the correction function can be written as a Fredholm integral equation of the second kind with a free term determined by f_w^c :

$$\begin{aligned} f_w^\Delta(x, \mathbf{k}) &= \int_{t_b}^0 dt \int d\mathbf{k}' f_w^\Delta(X(t), \mathbf{k}') S(\mathbf{k}', \mathbf{k}(t)) e^{-\int_t^0 \lambda(\mathbf{k}(\tau)) d\tau} + f_w^{\Delta,0}(x, \mathbf{k}) \\ f_w^{\Delta,0} &= \int_{t_b}^0 dt \left\{ \int d\mathbf{k}' f_w^c(X(t), \mathbf{k}') S(\mathbf{k}', \mathbf{k}(t)) e^{-\int_t^0 \lambda(\mathbf{k}(\tau)) d\tau} \right. \\ &\quad \left. - f_w^c(X(t), \mathbf{k}(t)) \lambda(\mathbf{k}(t)) e^{-\int_t^0 \lambda(\mathbf{k}(\tau)) d\tau} \right\} \end{aligned} \quad (5)$$

Here

$$X(t) = x - \int_t^0 \frac{\hbar K_x(\tau)}{m} d\tau \quad K_x(t) = k_x - \int_t^0 \frac{F(X(\tau))}{\hbar} d\tau \quad (6)$$

are classical Newton trajectories initialized by $x, k_x, 0, t < 0$, and $\mathbf{k}(t)$ stands for $K_x(t), \mathbf{k}_\perp$. The trajectory crosses the boundary of the device at a certain time t_b , where $f_w^\Delta(X(t_b), \mathbf{k}(t_b)) = 0$.

3 Computational Problem

The general task is to compute the averaged value of f_w^Δ in the given domain Ω of the two dimensional phase space. The averaged value can be expressed as:

$$I(\Omega) = \int dx \int dk_x f_w^\Delta(x, k_x) \theta_\Omega(x, k_x) = \int dx \int dk_x \int d\mathbf{k}_\perp f_w^\Delta(x, \mathbf{k}) \theta_\Omega(x, k_x) \quad (7)$$

by introducing the domain indicator $\theta_\Omega(x, k_x)$, which is unity if the arguments belong to Ω , and 0 otherwise. The solution of equation (5) can be expressed as consecutive iterations of the kernel on the free term: $f_w^\Delta = \sum_{p=0}^{\infty} f_w^{\Delta,p}$:

$$f_w^{\Delta,(p+1)} = \int_{-\infty}^0 dt \int d\mathbf{k}' \theta_D(X(t)) f_w^{\Delta,p}(X(t), \mathbf{k}') S(\mathbf{k}', \mathbf{k}(t)) e^{-\int_t^0 \lambda(\mathbf{k}(\tau)) d\tau} \quad (8)$$

The lower bound of the time integral has been extended to $-\infty$, since the introduced device domain indicator θ_D takes care for it's correct value t_b . We consider the contributions to (7) of the consecutive terms of (8). In this way we reduce the general task (7) to a problem of evaluation of the consecutive contributions:

$$I(\Omega) = \int dx \int dk_x \int d\mathbf{k}_\perp f_w^\Delta(x, \mathbf{k}) \theta_\Omega(x, k_x) = \sum_{p=0}^{\infty} \int d\mathbf{k}_\perp I_\Omega^{(p+1)}(\mathbf{k}_\perp)$$

$$I_\Omega^{(p+1)}(\mathbf{k}_\perp) = \int_{-\infty}^0 dt \int dx \int dk_x \int d\mathbf{k}' \theta_D(X(t))$$

$$f_w^{\Delta,p}(X(t), \mathbf{k}') S(\mathbf{k}', \mathbf{k}(t)) e^{-\int_t^0 \lambda(\mathbf{k}(\tau)) d\tau} \theta_\Omega(x, k_x) \quad (9)$$

The trajectory $X(t), \mathbf{k}(t) = (K_x(t), \mathbf{k}_\perp)$ is initialized by x, k_x at time 0, and the parameterization is backward: $t < 0$.

3.1 Stochastic Analysis

The aim of the following analysis is twofold: to devise a Monte Carlo method for evaluation of $I(\Omega)$; the method to be compatible with the established algorithms for device simulations and thus to allow an easy implementation. These algorithms emulate the natural processes of the evolution of Boltzmann carriers, which follow an incrementing in time succession. Thus equation (9) must be reformulated in a forward in time, $t > 0$, parameterization. According to (6) the trajectory is initialized by x, k_x at 0, which can be written as: $X(t) = X(t; x, k_x, 0) = x^t$ $K_x(t) = K_x(t; x, k_x, 0) = k_x^t$. Two basic properties of the Newton trajectories are utilized. A trajectory, being a unique solution of a first order differential equations, can be initialized by any of its points x^t, k_x^t associated to given time t . Furthermore, in stationary conditions trajectories are invariant with respect to a shift of both, the time origin and the parameterization time:

$$X(\tau) = X(\tau - t; x^t, k_x^t, 0) = X^t(\tau - t); \quad K_x(\tau) = K_x(\tau - t; x^t, k_x^t, 0) = K_x^t(\tau - t)$$

Here the initialization point/time have been changed accordingly, followed by a shift in time by $-t$. The short notations X^t, K^t recall for the novel initialization by $x^t, k_x^t, 0$. It follows that $x = X^t(-t)$, $k_x = K_x^t(-t)$. The Liouville theorem

$dxdk_x = dx^t dk_x^t$ is finally utilized to reformulate (9) as follows:

$$I_{\Omega}^{(p+1)}(\mathbf{k}_{\perp}) = \int_0^{\infty} dt \int dx^t \int dk_x^t \int d\mathbf{k}' \theta_D(x^t) f_w^{\Delta,p}(x^t, \mathbf{k}') \left\{ \frac{S(\mathbf{k}', k_x^t, \mathbf{k}_{\perp})}{\lambda(\mathbf{k}')} \right\} \\ \left\{ \lambda(K_x^t(t), \mathbf{k}_{\perp}) e^{-\int_0^t \lambda(K_x^t(\tau), \mathbf{k}_{\perp}) d\tau} \right\} \frac{\lambda(\mathbf{k}')}{\lambda(K_x^t(t), k_{\perp})} \theta_{\Omega}(X^t(t), K_x^t(t)) \quad (10)$$

where, now, the trajectory $X^t(t), K_x^t(t), t > 0$ is initialized by x^t, k_x^t at the time origin, and the equation has been augmented to obtain the (enclosed in curly brackets) well known Monte Carlo probability densities for scattering, $\mathcal{S}$, and drift, $\mathcal{D}$, processes. Indeed these densities associate to an initial point a final point within the scheme:

$$\mathcal{SD} \{x^t, \mathbf{k}' \rightarrow x^t, k_x^t, \mathbf{k}_{\perp} \Rightarrow X^t(t), K_x^t(t), \mathbf{k}_{\perp}\}, \quad (11)$$

where $\rightarrow$ corresponds to a scattering event, while $\Rightarrow$ to a drift, called also free flight. The scheme defines a segment of a numerical trajectory obtained by the consecutive iterations of (10). To analyze the physical aspects behind such a trajectory, it is sufficient to consider the second iteration $I_{\Omega}^{(2)}$. The following property will be used: in the limiting case, when the domain Ω shrinks to a point so that the domain indicator becomes a delta function: $\delta(x - X^t(t))\delta(k_x - K_x^t(t))$, equation (10) obtains a recursive form, due to the fact that $I_{\delta}^{(p+1)}(\mathbf{k}_{\perp}) = f_w^{\Delta,(p+1)}(x, k_x, \mathbf{k}_{\perp})$. A convention to mark the variables by the number of the corresponding iteration is followed, for convenience the superscript t is omitted along with the subscript of k_x . Finally, the notation (11), which provides a convenient abbreviation for the product of the two probability densities in (10) is utilized:

$$\int_0^{\infty} dt_2 \int dx_2 \int dk_2 \int d\mathbf{k}'_2 \theta_D(x_2) \int_0^{\infty} dt_1 \int dx_1 \int dk_1 \int d\mathbf{k}'_1 \theta_D(x_1) f_w^{\Delta,0}(x_1, \mathbf{k}'_1) \quad (12) \\ \mathcal{SD} \{x_1, \mathbf{k}'_1 \rightarrow x_1, k_1, \mathbf{k}'_{\perp 2} \Rightarrow X_1(t_1), K_1(t_1), \mathbf{k}'_{\perp 2}\} \frac{\lambda(\mathbf{k}'_1)}{\lambda(\mathbf{k}'_2)} \delta(x_2, k'_2; X_1 K_1, t_1) \\ \mathcal{SD} \{x_2, \mathbf{k}'_2 \rightarrow x_2, k_2, \mathbf{k}'_{\perp 3} \Rightarrow X_2(t_2), K_2(t_2), \mathbf{k}'_{\perp 3}\} \frac{\lambda(\mathbf{k}'_2)}{\lambda(\mathbf{k}'_3)} \theta_{\Omega}(X_2(t_2), K_2(t_2))$$

with

$$\delta(x_{s+1}, k'_{s+1}; X_s, K_s, t_s) = \delta(x_{s+1} - X_s(t_s))\delta(k'_{s+1} - K_s(t_s))$$

The zeroth order is given by the free term which, according to (5) has two components denoted by $f_w^{\Delta,0A}$ and $f_w^{\Delta,0B}$. The former is expressed in a forward in time parameterization [4] as follows:

$$f_w^{\Delta,0A}(x_1, \mathbf{k}'_1) = \int_0^{\infty} dt_0 \int dx_0 \int dk_0 \int d\mathbf{k}'_0 \theta_D(x_0) \left\{ \frac{\hbar^2 e^{-\frac{\hbar^2 \mathbf{k}'_0{}^2}{2mkT}}}{2\pi mkT} \right\} f_w^c(x_0, k'_0) \quad (13) \\ \left\{ \frac{S(\mathbf{k}'_0, k_0, \mathbf{k}'_{\perp 1})}{\lambda(\mathbf{k}'_0)} \right\} \left\{ \lambda(K_0(t_0), \mathbf{k}'_{\perp 1}) e^{-\int_0^{t_0} \lambda(K_0(\tau), \mathbf{k}'_{\perp 1}) d\tau} \right\} \frac{\lambda(\mathbf{k}'_0)}{\lambda(\mathbf{k}'_1)} \delta(x_1, k'_1; X_0 K_0, t_0)$$

The terms in the curly brackets in (12) and (13) correspond to a sequence of conditional probabilities giving rise to free-flight and scattering events. The final point of each free flight becomes the initial point for the next scattering event:

$$\begin{aligned}
 x_0, k'_0, \mathbf{k}'_{\perp 0} &\rightarrow x_0, k_0, \mathbf{k}'_{\perp 1} \Rightarrow X_0(t_0) = x_1, K_0(t_0) = k'_1, \mathbf{k}'_{\perp 1} & | & f_w^{\Delta, 0A}(x_1, \mathbf{k}'_1) \\
 x_1, k'_1, \mathbf{k}'_{\perp 1} &\rightarrow x_1, k_1, \mathbf{k}'_{\perp 2} \Rightarrow X_1(t_1) = x_2, K_1(t_1) = k'_2, \mathbf{k}'_{\perp 2} & | & f_w^{\Delta, 1A}(x_2, \mathbf{k}'_2) \\
 x_2, k'_2, \mathbf{k}'_{\perp 2} &\rightarrow x_2, k_2, \mathbf{k}'_{\perp 3} \Rightarrow X_2(t_2), K_2(t_2), \mathbf{k}'_{\perp 3} & | & I_{\Omega}^{(2)}(\mathbf{k}'_{\perp 3})
 \end{aligned}$$

The sequence of events resembles the evolution of a Boltzmann particle and thus enables the implementation of the standard algorithm for trajectory construction utilized in the device Monte Carlo simulators.

3.2 Numerical Aspects

We now return to the general task, the computation of $I(\Omega)$, and analyze what happens from a numerical point of view during the particle evolution. The basic notions from the Monte Carlo evaluation of integrals are assumed to be well known, and will be applied in the following. A general result is that a stochastic approach is optimal provided that the sampling probability density is proportional to the integrand function. In this respect the choice of the initial point $x_0, k'_0, \mathbf{k}'_{\perp 0}$ in (13) is according to the Gaussian in the first curly brackets for the transversal variables, and according to to:

$$\frac{|f_w^c(x_0, k'_0)|}{F_1}; \quad F_1 = \int dx \int dk_x |f_w^c(x, k_x)|;$$

for the longitudinal ones. Thus the initial weight of the particle is F_1 times the sign of f_w^c in the chosen point. The multiplication by F_1 can be done at the final stage of evaluation of the estimators, so that the initialized particle carries the sign only. The particle evolves to $x_1, k'_1, \mathbf{k}'_{\perp 1}$ as a result of a scattering and a drift event, and the weight is updated by the ratio of the two λ values. We note that at this stage the above procedure can be regarded as a legitimate experiment for evaluation of $I(\Omega)^{(0)} = \int d\mathbf{k}_{\perp 1} I_{\Omega}^{(0)}(\mathbf{k}_{\perp 1})$. An estimator $\xi_{\Omega}(0)$ is introduced, whose value is updated by adding of $\text{sign}(f_w^c)\lambda(\mathbf{k}'_0)/\lambda(\mathbf{k}'_1)$. The integral over the transverse variables means that the update of the estimator is independent of the concrete value of $\mathbf{k}_{\perp 1}$. The trajectory continues by a second scattering and free flight, and the weight is updated by the next fraction $\lambda(\mathbf{k}'_0)/\lambda(\mathbf{k}'_1)$. The obtained two-segment trajectory is a legitimate experiment for evaluation of $I(\Omega)^{(1)}$: the weight $\text{sign}(f_w^c)\lambda(\mathbf{k}'_0)\lambda(\mathbf{k}'_1)/\lambda(\mathbf{k}'_1)\lambda(\mathbf{k}'_2)$ is added to an estimator $\xi_{\Omega}(1)$. A third step follows in the same fashion, etc. The consecutive steps give rise to a weight $\text{sign}(f_w^c)\lambda(\mathbf{k}'_0)/\lambda(\mathbf{k}'_p)$ used to evaluate the consecutive values of $I(\Omega)^{(p)}$, stored by the corresponding estimators $\xi_{\Omega}(p)$. The procedure continues, until the trajectory abandons the device domain for the first time: In this case the

device domain indicator becomes zero, which resets the value of the accumulated weight of all further steps to 0. The contributions to the higher order terms in the sum for $I(\Omega)$ become zero and the further evolution of such a trajectory becomes obsolete. In this way one trajectory represents one independent experiment for a direct evaluation of I_Ω : all estimators can be merged into one, ξ_Ω . Finally, the arithmetic mean of the accumulated due to N independent trajectories value of ξ_Ω , multiplied by F_1 is a Monte Carlo estimate of I_Ω .

The contribution of the second component $f_w^{\Delta,0B}$ is a subject of similar analysis. The only difference is that the trajectory begins with a free flight, determined by the initialization point. This can be formally accounted by replacement of the first S/λ term in (13) by a delta function.

Different strategies may be considered: the two contributions can be evaluated separately, or $f_w^{\Delta,0}$ can be evaluated at a first stage and then used for a direct evaluation of the iteration series. As the efficiency of these strategies can be estimated by numerical experiments only, we continue by adopting the 'separate simulation' approach.

3.3 Pointwise Evaluation

It is further assumed that the coherent solution is known only pointwise. The following decomposition can be utilized in (10):

$$\int dx^t \int dk'_x f_w^{\Delta,(p)}(x^t, k'_x, \mathbf{k}_\perp) = \sum_{mn} f_w^{\Delta,(p)}(x_m^t, k'_{xn}, \mathbf{k}_\perp) \Delta \quad (14)$$

introduced by the interval $\Delta = \Delta_{kx} \Delta_x$. The computational task is further focused on the evaluation of the averaged value of $f_w^{\Delta,(p+1)}$ in the domain Ω_{ij} specified by Δ around (x_i, k_{xj}) . In particular (10) reduces to the recursive relation:

$$f_w^{\Delta,(p+1)}(x_i, k_{xj}, \mathbf{k}_\perp) = \sum_{mn} \int dk_x^t \int_0^\infty dt \int d\mathbf{k}_\perp f_w^{\Delta,p}(x_m^t, k'_{xn}, \mathbf{k}'_\perp) \left\{ \frac{S(\mathbf{k}', k_x^t, \mathbf{k}_\perp)}{\lambda(\mathbf{k}')} \right\} \\ \left\{ \lambda(K_x^t(t), \mathbf{k}_\perp) e^{-\int_0^t \lambda(K_x^t(\tau), \mathbf{k}_\perp) d\tau} \right\} \frac{\lambda(\mathbf{k}')}{\lambda(K_x^t(t), \mathbf{k}_\perp)} \theta_D(x_m^t) \theta_{\Omega_{ij}}(X^t(t), K_x^t(t)), \quad (15)$$

where the trajectory is initialized by x_m, k_x^t , and gives rise to the following algorithm:

- The phase space simulation domain is decomposed into sub-domains Ω_{mn} around x_m, k_{xn} nodes; The estimators ξ_{mn} are initialized to zero. Evaluated are the probabilities:

$$P_{mn} = \frac{|f_w^c(x_m, k_{xm})|}{F_1}; \quad F_1 = \sum_{mn} |f_w^c(x_m, k_{xm})|;$$

The number of independent Monte Carlo experiments is specified to N_l .

- Within a loop over $l = 1, \dots, N_l$: the initial point $x_m, k_{xn}, \mathbf{k}_\perp$ of the l -th trajectory is chosen randomly by using P_{mn} and the Gaussian distribution function of the transversal wave vectors. The product of the sign of f_w^c and λ , both evaluated at the initial point, is assigned to a variable w_l .
- The construction of the trajectory begins by a scattering event for the iteration series A corresponding the first component of the free term, followed by a free flight. For the second component, B , only the free flight remains. In both cases the events are realized by the standard scheme for device Monte Carlo simulators.
- After each free flight: if the trajectory belongs to the device domain, the estimator of the nearest to the end point node is updated by adding w_l/λ where λ is determined by the free flight end point; otherwise the construction of the trajectory is stopped and another trajectory begins.
- At the end of the loop the values of the estimators are divided by N_l . It holds:

$$f_w^{\Delta A, B}(x_i, k_{xj}) \simeq \xi_{ij}^{A, B}/N_l.$$

Finally

$$f_w^\Delta(x_i, k_{xj}) = f_w^{\Delta A}(x_i, k_{xj}) - f_w^{\Delta B}(x_i, k_{xj}).$$

4 Conclusions

The presented approach aims at an estimation of the effect of scattering to the coherent transport in nanoscale devices. It offers high computational efficiency at the expense of neglecting the correlations between electrical potential and scattering events. The devised Monte Carlo algorithm calculates pointwise the values of the scattering-induced Wigner function correction. It is compatible with the established methods for Monte Carlo device simulations and thus allows an easy implementation.

Acknowledgment

This work has been supported by the Austrian Science Fund Project FWF-P21685.

References

1. Querlioz, D., Dollfus, P.: The Wigner Monte Carlo Method for Nanoelectronic Devices - A particle description of quantum transport and decoherence (ISTE-Wiley) (2010)
2. Kosina, H., Nedjalkov, M., Selberherr, S.: The stationary Monte Carlo method for device simulation - Part I: Theory. J. Appl. Phys. 93(6), 3553–3563 (2003)
3. Nedjalkov, M., Kosina, H., Selberherr, S., Ringhofer, C., Ferry, D.K.: Unified particle approach to Wigner-Boltzmann transport in small semiconductor devices. Physical Review B 70(11), 115319–115335 (2004)
4. Schwaha, P., Baumgartner, O., Heinzl, R., Nedjalkov, M., Selberherr, S., Dimov, I.: Classical approximation of the scattering induced Wigner correction equation. In: 13th International Workshop on Computational Electronics Book of Abstracts, IWCE-13, Beijing, China, pp. 177–180. IEEE, Los Alamitos (2009)

Modeling Thermal Effects in Fully-Depleted SOI Devices with Arbitrary Crystallographic Orientation

K. Raleva¹, D. Vasileska², and S.M. Goodnick²

¹ University Sts, Cyril and Methodius, Skopje, Republic of Macedonia

² Arizona State University, Tempe, AZ 85287-5706, USA

Abstract. In this work we continue our investigation on the heating effects in nano-scale FD-SOI devices using an in-house thermal particle-based device simulator. We focus on the current variations for FD-SOI devices with arbitrary crystallographic orientation and examine which crystallographic orientation gives better results from electrical and thermal point of view. Our simulation results demonstrate that one can obtain the lowest current degradation with (110) wafer orientation. The temperature of the hot-spot is the smallest for (110)-orientation as well.

Keywords: nano-scale FD-SOI devices, self-heating effects, crystallographic orientation, particle-based device simulations.

1 Introduction

The continuous downscaling of MOSFET geometries is motivated by the need for higher packing density and device speed. The objective of the device miniaturization is to deliver high performance at low costs. It results in reduced unit cost per function and in enhanced performance. Full functionality in MOSFETs with technological gate lengths between 10 nm and 100 nm has been achieved leading to mass production of devices, and MOSFETs below 10 nm gate lengths have been established. Maintaining the pace of MOSFET device scaling in the sub-100 nm gate length regime has become increasingly difficult. The simple scaling of the channel length and gate oxide thickness is no longer sufficient to deliver the projected speed/power performance enhancement for high performance logic device technologies. Problems include short-channel effects such as, subthreshold leakage current and threshold voltage changes due to the drain-induced barrier lowering (DIBL), and the high level of leakage current through the ultra-thin gate dielectric. These leakage currents cause higher static power dissipation. Active switching power is another key problem where a higher number of gates switching at high frequency with only modest reductions in supply voltage result in high active power density. The problems facing device scaling necessitate new solutions. The desired solution is one that increases MOSFET drive current while reducing leakage currents, short-channel effects and the active power density. To achieve further improvement of performance in scaled silicon devices

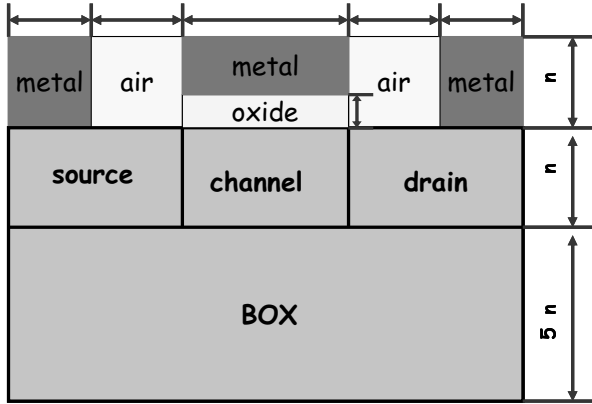


Fig. 1. Cross-section and geometrical dimensions of the simulated 25nm gate-length FD-SOI structure

applied mechanical stress [1], alternative wafer orientations [2],[3], and multi-gate transistors [4], [5] have been actively researched or are already in production. All these options take advantage of the anisotropic nature of the silicon crystal, and therefore, of its anisotropic bandstructure; in engineering terms gains are utilized in the carrier transport mass and mobility. For instance, strained Si is the only new channel material which has recently made its way into the commercial integrated circuits. By straining the silicon channel, carrier mobility can be enhanced. Also, devices fabricated on strained-Si (110) wafer orientations has shown improved mobility characteristics over (100) devices [6]. Similar results for (110) strained SOI MOSFETs have been published in Ref. [7] as well.

The current trend in device scaling is a transition away from conventional planar CMOS to alternative non-planar technology devices, such as fully-depleted (FD), dual-gate (DG), tri-gate silicon-on-insulator (SOI) and others. The advantages of these devices are higher drive current, low junction capacitance, reduced leakage current, suppression of floating-body effects, absence of latch-up and ease in scaling. But, one of the major problem with SOI devices is that they exhibit self-heating effects. These self-heating effects arise from the fact that the underlying SiO_2 layer has about 100 times smaller thermal conductivity than bulk Si.

We have previously reported that self-heating and increased power density play important roles in the operation of FD-SOI devices with gate lengths between 25 and 180 nm [8,9,10] using 2D electro-thermal simulation based on the self-consistent solution of the Boltzmann transport equation for the electrons via Monte Carlo techniques and the energy balance equations for acoustic and optical phonons. There, it was shown that due to geometry and velocity overshoot, self-heating effects are more pronounced for larger channel length devices with correspondingly larger supply voltages.

In this work we continue our investigation on the current degradation in nano-scale FD-SOI devices due to self-heating effects. We focus on the current

variations for FD-SOI devices with arbitrary crystallographic orientation and examine which crystallographic orientation gives better results from electrical and thermal point of view. Details of the structure being examined, electrical and thermal boundary conditions, current degradation due to self-heating and lattice temperature profiles for (100), (111) and (110) wafer orientations are presented in Section 2 of this paper. Conclusions regarding this work and future directions of research are given in Section 3.

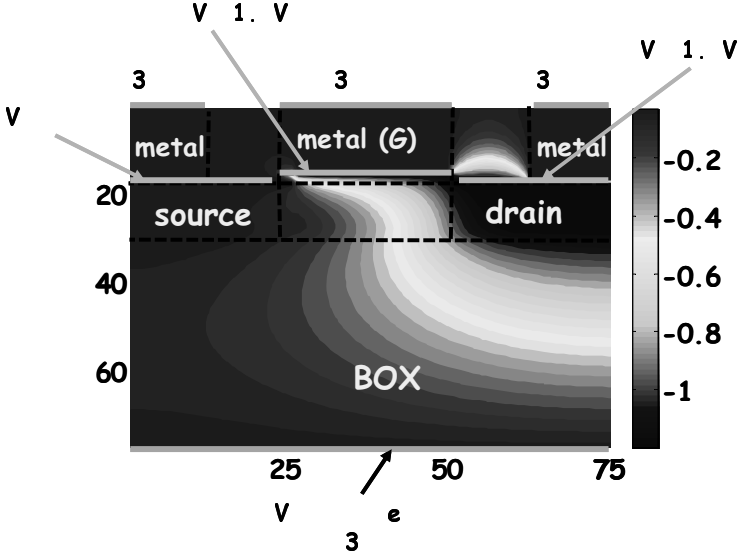


Fig. 2. Conduction band edge profile (in Volts) of the simulated structure for $V_{GS} = 1.2V$ and $V_{DS} = 1.2V$. Also, the positions of the thermal Dirichlet boundary conditions are shown.

2 Electro -Thermal Simulations for 25 nm Gate-Length FD-SOI MOSFET

The cross-section of the simulated 25 nm gate-length FD-SOI structure is shown in Fig. 1. In order to get more realistic results from thermal simulations, we extend the length of the metal (copper) gate, source and drain electrodes. In all simulation presented in this work, we have assumed Dirichlet boundary conditions at the bottom of the BOX and at the end of the three electrodes (see Fig. 2). For all other boundaries, Neumann conditions are assumed. Details on the role of the substrate and thermal boundary conditions can be found in [9]. The conduction band edge profile for $V_{GS} = 1.2V$ and $V_{DS} = 1.2V$ and the electric Dirichlet boundary conditions are also shown in Fig. 2.

To take into account the wafer orientation, we use the standard effective mass approach which describes the band edge electronic properties in an approximate manner. Silicon -valley effective masses and subband degeneracy for (100), (111)

Table 1. Silicon Δ -valley effective masses and subband degeneracy for (100), (111) and (110) wafer orientations. ($m_l=0.91$, $m_t=0.19$)

Wafer	m_x (transport eff. mass)	m_y (confinement eff. mass)	m_z (width eff. mass)	Deg .
(100)	m_t	m_l	m_t	2
	m_l	m_t	m_t	2
	m_t	m_t	m_l	2
(111)	$3m_l m_t / (2m_l + m_t)$	$(2m_l + m_t) / 3$	m_t	2
	$(2m_l + m_t) / 6$	$3m_t (m_l + m_t) / (2m_l + m_t)$	$2m_l m_t / (m_l + m_t)$	4
(110)	m_t	$2m_l m_t / (m_l + m_t)$	$(m_l + m_t) / 2$	4
	m_l	m_t	m_t	2

and (110) wafer orientations are given in Table 1, where m_l and m_t are the longitudinal and the transverse effective masses, respectively. The expressions for the effective mass are derived according to [11]. In Table 2 we present the on-current variations and current degradations due to self-heating for different wafer orientation.

Table 2. Current variations for different wafer orientations

Wafer orientation	# iteration	Current (mA/ μ m)	Current Degradation (%)
(100)	isotherm.	1.828	\
(100)	thermal	1.757	3.88
(111)	isotherm.	1.748	\
(111)	thermal	1.655	5.32
(110)	isotherm.	1.825	\
(110)	thermal	1.785	2.19

The simulation results show that the higher value of the on-current is obtained when the simulated FD-SOI structure is designed on wafers with either (100) or (110) crystallographic orientations which is due to the lower effective masses along the corresponding transport directions which results in a higher electron drift velocity in the channel (see Fig. 3). Note that the carriers in the simulated structures for the given bias conditions are in the velocity overshoot regime which leads to very small current degradation.

The lattice temperature profiles in the active silicon layer for (100) and (111) crystallographic orientations are shown in Fig. 4 (left panel). From Fig. 4 (right

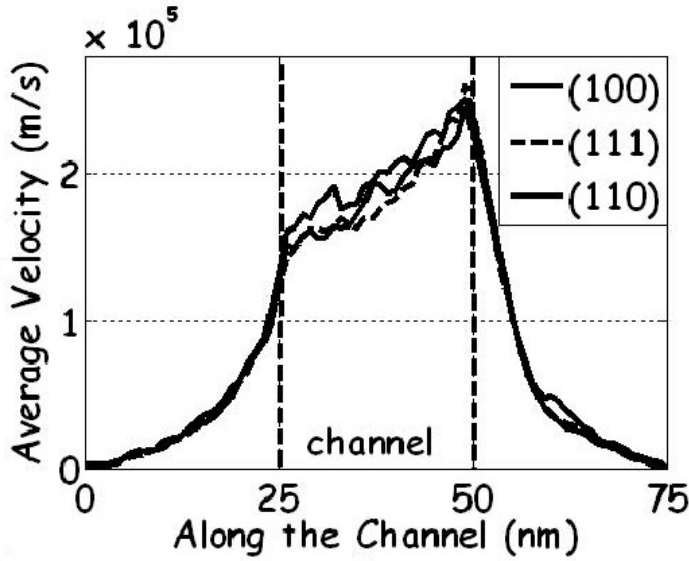


Fig. 3. Average electron velocity along the channel for different wafer orientations

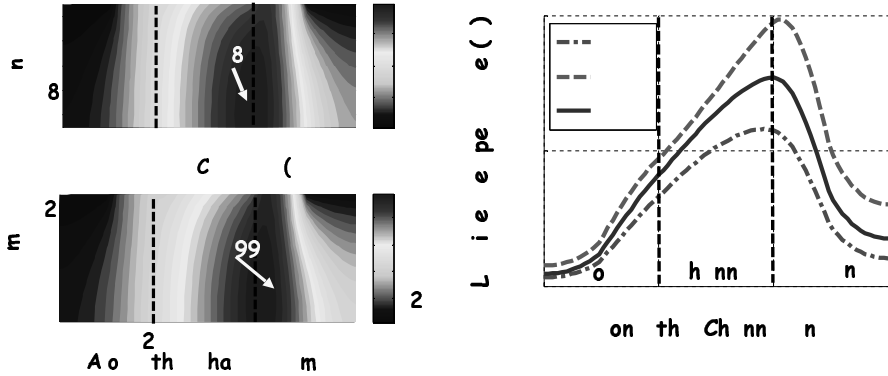


Fig. 4. Left Panel: Lattice temperature profile in the active Si-layer for (100) (top) and (111) (bottom) wafer orientation. Right Panel: Average lattice (acoustic phonon) temperature profile along the channel in the active Si-layer for different wafer orientations.

panel), one can observe that the position of the hot-spot region does not change with the wafer orientation, but the maximum temperature of the hot-spot is highest for (111) wafer orientation. The higher lattice temperature reduces the thermal conductivity in the channel as can be seen in Fig. 5. These results are obtained by using our novel theoretical model for the temperature and thickness dependence of the thermal conductivity [12], which is derived for (100) wafer

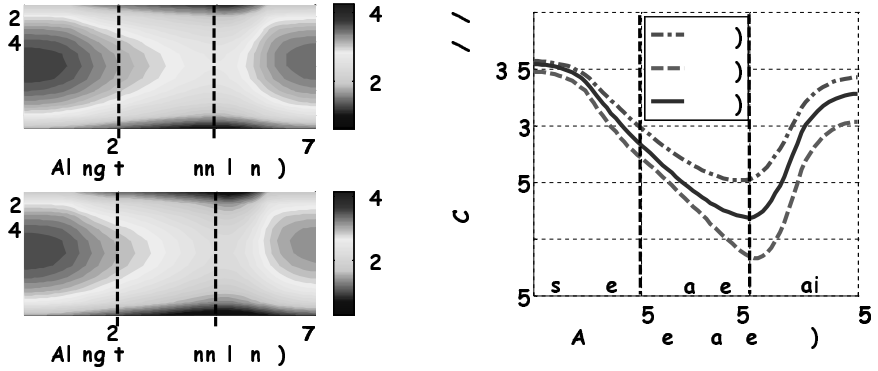


Fig. 5. Left Panel: Thermal conductivity profile in the active Si-layer for (100) (top) and (111) (bottom) wafer orientation. Right Panel: Average thermal conductivity profile along the channel in the active Si-layer for different wafer orientations.

orientation. We believe that the inclusion of the proper thermal conductivity model for (110) and (111) wafer orientation will decrease the current degradation even more. The results of these simulations will be presented at the conference.

3 Conclusions and Future Works

In this work we have presented preliminary simulation results for self-heating effects in nanoscale FD-SOI devices with arbitrary crystallographic orientations. The results from this work are consistent with the results given in [3]. The simulation results for three different wafer orientations ((100), (111) and (100)) show that the velocity overshoot leads to insignificant current degradation because of the self-heating effects. The main result of our analysis is that one can obtain the lowest current degradation with (110) orientation. There are many more issues that need to be addressed to derive even more conclusive results regarding self-heating in non-(100) FD-SOI structures such as the inclusion of the temperature and position dependence of the thermal conductivity for corresponding wafer orientation, or the inclusion of thermal conductivity tensors for arbitrary crystallographic orientations. This work is currently underway and will be presented elsewhere.

References

1. Mistry, K., et al.: Delaying forever: Uniaxial strained silicon transistors in a 90nm CMOS technology. In: 2004 Symposium on VLSI Technology, Digest of Technical Papers, pp. 50–51 (June 2004)
2. Yang, M., et al.: Performance Dependence of CMOS on Silicon Substrate Orientation for Ultrathin Oxynitride and HfO₂ Gate Dielectrics. IEEE Transactions on Electron Devices 51(10), 1621–1626 (2004)

3. Chang, L., Jeong, M., Yuang, M.: CMOS Circuit Performance Enhancement by Surface Orientation Optimization. *IEEE Transactions on Electron Devices* 55(6), 1306–1316 (2008)
4. Wong, H.-S.P.: Beyond the Conventional Transistor. *IBM J. Res. Dev.* 46(2/3), 133–168 (2002)
5. Chau, R.S.: Integrated CMOS Tri-Gate Transistors: Paving the Way to Future Technology Generations. *Technology@Intel Magazine*, pp. 1–7 (August 2006)
6. Mizuno, T., Sugiyama, N., Tezuka, T., Takagi, S.: (110) strained-SOI n-MOSFETs with higher electron mobility. *IEEE Electron Device Letters* 24(4), 266–268 (2003)
7. Mizuno, T., Sugiyama, N., Tezuka, T., Moriyama, Y., Nakaharai, S., Takagi, S. (110)-surface strained-SOI CMOS devices. *IEEE Transactions on Electron Devices* 52(3), 367–734 (2005)
8. Raleva, K., Vasileska, D., Goodnick, S.M., Nedjalkov, M.: Modeling Thermal Effects in Nano-devices. *IEEE Transactions on Electron Devices* 55(6), 1306–1316 (2008)
9. Vasileska, D., Raleva, K., Goodnick, S.M.: Self-Heating Effects in Nano-Scale FD SOI Devices: The Role of the Substrate, Boundary Conditions at Various Interfaces and the Dielectric Material Type for the BOX. *IEEE Transactions on Electron Devices* 56(12), 3064–3071 (2009)
10. Raleva, K., Vasileska, D., Goodnick, S.M.: Is SOD Technology the Solution to Heating Problems in SOI Devices? *IEEE Electron Device Letters* 29(6), 621–624 (2008)
11. Rahman, A.: Exploring new channel materials for nanoscale CMOS devices: a simulation approach, PhD Thesis (Purdue University) (December 2005)
12. Vasileska, D., Raleva, K., Goodnick, S.M.: Electrothermal Studies of FD SOI Devices That Utilize a New Theoretical Model for the Temperature and Thickness Dependence of the Thermal Conductivity. *IEEE Transactions on Electron Devices* 57, 726–728 (2010)

Particle Monte Carlo Algorithms with Small Number of Particles in Grid Cells

Stefan K. Stefanov

Institute of Mechanics, Bulgarian Academy of Sciences,
1113 Sofia, Bulgaria

stefanov@imbm.bas.bg

<http://www.imbm.bas.bg/index.php?page=157>

Abstract. The Direct Simulation Monte Carlo (DSMC) analysis of two- and three-dimensional rarefied gas flows requires computational resources of very large proportions. One of the major causes for this is that, along with the multidimensional computational mesh, the standard DSMC approach also requires a large number of particles in each cell of the mesh in order to obtain sufficiently accurate results. In this paper we present two modified simulation procedures which allow more accurate calculations with a smaller mean number of particles ($\langle N \rangle \sim 1$) in the grid cells. In the general DSMC scheme, the standard DSMC collision algorithm is replaced by a new collision procedure based on "Bernoulli trials" scheme or its simplified version. The modified algorithms use a symmetric Strang splitting scheme that improves the accuracy of the splitting method to $O(\tau^2)$ with respect to the time step τ making the modified DSMC method a more effective numerical tool for both steady and unsteady gas flow calculations on fine multidimensional grids. Here the considered modifications are validated on the one-dimensional unsteady-state problem of strong shock wave formation.

Keywords: Direct Simulation Monte Carlo (DSMC) method, kinetic theory, rarefied gas flow, micro gas flow.

1 Introduction

The Direct Simulation Monte Carlo (DSMC) technique [1] is a powerful numerical method for studying rarefied gas dynamics and micro gas flow problems. The DSMC technique uses a finite set of model particles denoted by their positions and velocities $\{\mathbf{x}_i, \xi_i\}$, $i = 1, \dots, N$, that move and collide in a computational domain to perform a stochastic simulation of the real molecular gas dynamics. The basic concept of the method is built on a discretization in time and space of the real gas dynamics process and splitting the motion into two successive stages of free molecular motion and binary intermolecular collisions within the grid cells each time step. The second stage of modeling the binary collisions in cells is more complicated and over the years serious efforts have been made to improve the "Time Counter" collision scheme originally proposed by Graham

Bird [2]. Later, as a result of subsequent theoretical investigations, several collision schemes with better characteristics have been proposed: "Null-Collision" [3], "Ballot-Box" [4], "Modified-Nambu" [5], "Majorant Collision Frequency" [6], and "No Time Counter (NTC)" [1]. The most frequently used scheme has become the Bird's NTC scheme and further in the text we will refer to it as the "standard scheme". In the main, all these schemes require a large number of particles per cell ($N \sim 10 - 20$) in order to obtain reliable results. The reason is that all these algorithms allow multiple repeated collisions between one and the same particle pair that lead to distortion of the collision process in cells with small number of particles.

In the present paper two modified collision algorithms that avoid the generation of repeated collisions in cells are proposed to replace the standard NTC collision procedure in the case of small mean number of particles ($N \sim 1, 2$) in cells. The first is the so-called Bernoulli trials (BT) scheme originally proposed by Yanitskiy [4]. However, the BT algorithm is computationally more intensive with respect to the number of particles per cell than the standard NTC one. The computational intensity of the BT algorithm is proportional to $O(N^2)$ while NTC one is proportional to $O(N)$. The efficiency [7] of both algorithms becomes almost equivalent when the number of particles is dropped to 1, 2. The second algorithm is a simplification of the Yanitskiy's Bernoulli trials scheme with a decreased number of computations per cell of order $O(N)$. Considered from viewpoint of the general simulation algorithm, each of these algorithms can replace the standard NTC collision procedure in a two-step collision procedure, which represents an intrinsic part of a symmetric Strang splitting scheme [8]. The two-collision scheme has been applied successfully for the simulation of the three-dimensional Rayleigh-Bénard convection of a rarefied gas [9]. In this paper some validation results for the simplified BT algorithm are presented, which are obtained from the simulation of the one-dimensional shock wave formation in front of a moving supersonic speed piston [1].

2 The Modified Collision Algorithms

A detailed mathematical description of the motion of a rarefied gas system can be given by an evolutionary kinetic equation in the following non-closed form with respect to the velocity distribution function $f(t, \mathbf{x}, \xi)$:

$$\frac{\partial}{\partial t} f(t, \mathbf{x}, \xi) = -D[f(t, \mathbf{x}, \xi)] + Q[f^{(2)}(t, \mathbf{x}, \xi, \mathbf{x}_*, \xi_*)], \quad (1)$$

where $f(t, \mathbf{x}, \xi) = f^{(1)}(t, \mathbf{x}, \xi)$ and $f^{(2)}(t, \mathbf{x}, \xi, \mathbf{x}_*, \xi_*)$ are one-particle and two-particle distribution functions of the particle velocities ξ and ξ_* at time t and spacial coordinate $\mathbf{x}$, D denotes a linear differential operator describing the free particle motion and Q is a non-linear integral operator describing the particle binary interactions. For more details concerning equation (1) and its relation to the Boltzmann equation we refer to Cercignani's monograph [10]. We denote by operators $S_Q^{\tau, h}$ and S_D^τ the numerical algorithms approximating the action of the

collision and convective terms in Eq. (1), respectively. If we denote by $S_{Q+D}^{\tau,h}$ the operator evaluating the solution of (1) at $t_k + 1$ from the state at t_k then the splitting method is expressed with the approximation

$$S_{Q+D}^{\tau,h} \approx S_D^\tau S_Q^{\tau,h}. \quad (2)$$

Using the result, obtained by Bobylev and Ohwada [11], one can show that the splitting method approximates the Boltzmann equation with accuracy $O(\tau + h)$. The accuracy with respect to time step can be improved by using the Strang splitting symmetric scheme [8]:

$$S_{Q+D}^\tau(f_0) = S_Q^{\tau/2} \left\{ S_D^\tau \left[S_Q^{\tau/2}(f_0) \right] \right\} + O(\tau^3). \quad (3)$$

Further, the considered collision algorithms are presented by operator $S_Q^{\tau,h}$. A detailed description of the general standard two-stage DSMC algorithm can be found in the Bird's monograph [1]. The problems related to the algorithm convergence to the Boltzmann equation solution are considered by Wagner [12]. During a particle simulation the following two stages are performed over each time step (t_k, t_{k+1}) , $k = 1, \dots, K$:

Stage 1. Operator $S_Q^{\tau,h}$ (The standard NTC collision procedure). Three steps are included in the “No Time Counter” collision procedure performed in each cell l , $l = 1, \dots, M$:

- computing the number of particle pairs N_c to be checked for a collision;
- acceptance-rejection of each pair (i, j) , $1 \leq i < j \leq N^{(l)}$, chosen at random from the particle subset $N^{(l)}$;
- if the collision is accepted then the particle velocities are changed to their post-collision values. During stage 1, the particle positions are not changed.

Stage 2. Operator S_D^τ (Free particle motion). Each particle $\mathbf{x}_i, \xi_i$, $i = 1, \dots, N$ is moved over the time step τ to its new position $\mathbf{x}'_i = \mathbf{x}_i + \xi_i \tau$. The boundary conditions are also simulated within Stage 2.

The required number of operations in a cell l is $O(N^{(l)})$. The standard NTC collision algorithm allows multiple repeated collisions of one and the same particle pair. As a consequence, the major effect on a standard DSMC simulation with small number of particles in cells is a reduction in the local collision frequency, which only converges to the Boltzmann collision frequency for large enough number of particles per cell. The stochastic properties of the collision algorithm can be improved if the standard NTC algorithm is replaced by a collision algorithm using Bernoulli trials or its simplifications. In order to derive the Bernoulli trials scheme we will follow Yanitskiy [4]. It is known that in the case of binary collisions equation (1) without the convection term can be described by the famous stochastic model of Kac [13]. Consider the evolution of a particle system $\{\mathbf{x}^{(l)}, \Xi_{N^{(l)}}\} = \{\mathbf{x}_j^{(l)}(t_k), \xi_j^{(l)}(t_k)\}$, $j = 1, \dots, N^{(l)}$ in cell (l) for time τ . The Kac stochastic model can be described by the following set of postulates:

- time intervals $\delta t_m = t_m - t_{m-1}$ between two binary collisions $m - 1$ and m are distributed according to exponential law

$$Prob \{ \delta t > t \} = e^{-\nu t}, \quad (4)$$

where

$$\nu = \sum_{1 \leq i < j \leq N} w_{ij}, \quad w_{ij} = \frac{\sigma_{ij} g_{ij}}{V^{(l)}}, \quad (5)$$

- number of collisions $s(\tau)$ within time τ is defined by the condition

$$\sum_{m=1}^s \delta t_m \leq \tau < \sum_{m=1}^{s+1} \delta t_m; \quad (6)$$

- probability for collision of pair (i, j) is equal to

$$W_{ij}^* = \frac{w_{ij}}{\nu}; \quad (7)$$

- if collision is accepted, velocities (ξ_i, ξ_j) are changed to their post-collision values (ξ'_i, ξ'_j) . If collision is rejected velocities remain unchanged.

Presented in operator form the Kac master equation reads as follows [4]

$$\begin{aligned} \frac{\partial}{\partial t} F_{N^{(l)}}(t, \mathbf{x}^{(l)}, \Xi_{N^{(l)}}) = \\ = \left[\sum_{1 \leq i < j \leq N^{(l)}} w_{ij} (T_{ij} - I) \right] F_{N^{(l)}}(t, \mathbf{x}^{(l)}, \Xi_{N^{(l)}}) = \nu (T - I) F_{N^{(l)}}(t, \mathbf{x}^{(l)}, \Xi_{N^{(l)}}), \end{aligned} \quad (8)$$

where

$$\begin{aligned} I\psi &\equiv \psi, \quad T_{ij} = \int_{4\pi} \psi(\Xi_{ij}) B(g_{ij}, \theta) d\Omega(\theta), \\ T\psi &= \sum_{1 \leq i < j \leq N^{(l)}} w_{ij} T_{ij} \psi \end{aligned} \quad (9)$$

are operators, acting on a linear normal space of continues functions $\psi(\Xi)$ over Ω , function $B(g_{ij}, \theta)$ is a scattering kernel (see [10]). The solution at time t in the case of a given state at previous time t_0 can be expressed as follows

$$F_{N^{(l)}}(t, \mathbf{x}^{(l)}, \Xi_{N^{(l)}}) = G(t) F_{N^{(l)}}(t_0, \mathbf{x}^{(l)}, \Xi_{N^{(l)}}), \quad (10)$$

where the transition operator $G(t)$ takes the form

$$G(t) = \exp \left[t \sum_{1 \leq i < j \leq N^{(l)}} w_{ij} (T_{ij} - I) \right] = \exp [t (T - I)]. \quad (11)$$

For small interval τ one can obtain for G

$$G(\tau) = \prod_{1 \leq i < j \leq N^{(l)}} e^{\tau w_{ij} (T_{ij} - I)} = \prod_{i=1}^{N^{(l)}-1} \prod_{j=i+1}^{N^{(l)}} e^{\tau w_{ij} (T_{ij} - I)}. \quad (12)$$

If each term is extended in a series of τ and the terms of order equal or higher than $O(\tau^2)$ are neglected the equation (12) is reduced to

$$\begin{aligned} G_1(\tau) &= \prod_{i=1}^{N^{(l)}-1} \prod_{j=i+1}^{N^{(l)}} [(1 - \tau w_{ij}) I + \tau w_{ij} T_{ij}] \\ &= \prod_{i=1}^{N^{(l)}-1} \prod_{j=i+1}^{N^{(l)}} [(1 - W_{ij}) I + W_{ij} T_{ij}]. \end{aligned} \quad (13)$$

The variable W_{ij} in the right hand side of equation (13) is a probability for collision of the particle pair (i, j) within time τ in the proposed by Yanitskiy [4] Bernoulli trials BT collision algorithm (for brevity - **algorithm A**).

Algorithm A:

- For each cell l ($l = 1, \dots, M$) (with volume $V^{(l)}$:
- each pair of particles in cell l with velocities $(\xi_i, \xi_j), i < j = 1, \dots, N^{(l)}$ is checked for collision with probability

$$W_{ij} = \frac{\sigma_{ij} g_{ij} \tau}{V^{(l)}}, \quad (14)$$

- where probability W_{ij} must satisfy the condition $Prob\{W_{ij} < 1\} \rightarrow 0$.
- if collision is accepted then velocities (ξ_i, ξ_j) are changed to the post-collision values (ξ'_i, ξ'_j) , otherwise they remain unchanged.

The **algorithm A** requires number of operations $O(N^{(l)^2})$ in cell l . The square dependance can be compensated when the number of particles in cell is small $N^{(l)} \rightarrow 1$. Here we propose a simplification of the BT algorithm, which realizes a number of computations proportional to the number of particles $N^{(l)}$. The new modification follows from the simplification of the transition operator (13). If we extend the internal product in the right hand side of (13) in a series of j with respect to τ and restrict ourself by the terms of order $O(\tau)$ then we get a new transition operator $G_2(\tau)$

$$G_2(\tau) = \prod_{i=1}^{N^{(l)}-1} \left[\left(1 - \sum_{j=i+1}^{N^{(l)}} \tau w_{ij} \right) I + \sum_{j=i+1}^{N^{(l)}} \tau w_{ij} T_{ij} \right]. \quad (15)$$

If $k = (N^{(l)} - i)$ and term τw_{ij} is replaced by $k^{-1}((k\tau w_{ij}))$ everywhere in equation (15) one obtains

$$G_2(\tau) = \prod_{i=1}^{N^{(l)}-1} \left[\left(1 - \sum_{j=i+1}^{N^{(l)}} \frac{1}{k} (k\tau w_{ij}) \right) I + \sum_{j=i+1}^{N^{(l)}} \frac{1}{k} ((k\tau w_{ij}) T_{ij}) \right]. \quad (16)$$

The algorithmic interpretation of operator (16) presents a new algorithm with linear computational intensity with respect to $N^{(l)}$ for simulation of binary collisions with time τ . Its description is given as:

Algorithm B:

- a sequence of pairs $i = 1, \dots, (N^{(l)} - 1)$ is chosen from $N^{(l)}$ particles in cell l as follows:
 - the first particle i is the particle with index i in the particle list created for cell l ;
 - the second particle $j \in [i + 1, N^{(l)}]$ is chosen with probability $1/k$ from $k = (N^{(l)} - i)$ particles taking place in the list after particle i .
- particle pair (i, j) is checked for collision with probability

$$\hat{W}_{ij} = k \frac{\sigma_{ij} g_{ij} \tau}{V^{(l)}}, \quad (17)$$

where probability $\hat{W}_{ij}$ must satisfy the condition

$$Prob\{\hat{W}_{ij} \geq 1\} \rightarrow 0; \quad (18)$$

- if collision is accepted then velocities (ξ_i, ξ_j) are changed to the post-collision values (ξ'_i, ξ'_j) , otherwise they remain unchanged.

It is worth noting that the better efficiency of algorithm B with respect to the number of particles in cell requires a stronger limitation of the time step, i.e. a smaller time step (see condition (18)). At the same time both algorithms, A and B, avoid the realization of repeated collisions within a time step τ . In the framework of the general Strang splitting scheme algorithms A or B are applied twice with step $\tau/2$ realizing of the action of operator $S_Q^{(\tau/2, h)}$.

3 Numerical Validation of Algorithms A and B

In order to validate numerically the proposed modification in the DSMC method we consider the one-dimensional unsteady-state problem of the formation of a strong shock wave by a piston that impulsively starts to move with a constant velocity of 2285.5 m/s in the x-direction. The gas in front of the piston uses the variable hard sphere model of argon at a temperature of $T_0 = 273^\circ\text{K}$ and a number density of $n_0 = 10^{20} \text{ m}^{-3}$. The piston can travel within an interval of 1 m i.e. approximately 100 mfps in the undisturbed gas. This is exactly the same case as described in Bird's book [1](§13.2). Bird's original program DSMC1U has been used to obtain the results by the standard DSMC method. We have used the same source code to obtain the modified scheme, substituting the "No Time Counter" collision procedure by the Bernoulli trials one (algorithm A) or its simplified version (algorithm B). The results are normalized by the molecular mean free path for length, the most probable molecular speed $V_{th} = \sqrt{2RT_0}$ for velocity and T_0 for temperature. Figure 1 presents the profiles of x-velocity (a) and temperature (b) along the x-direction, computed by both the standard NTC algorithm and simplified BT algorithm B on a grid with 400 uniform cells with different number of particles per cell. The reference profiles (shown as thick solid lines) are computed by the standard "No Time Counter" NTC

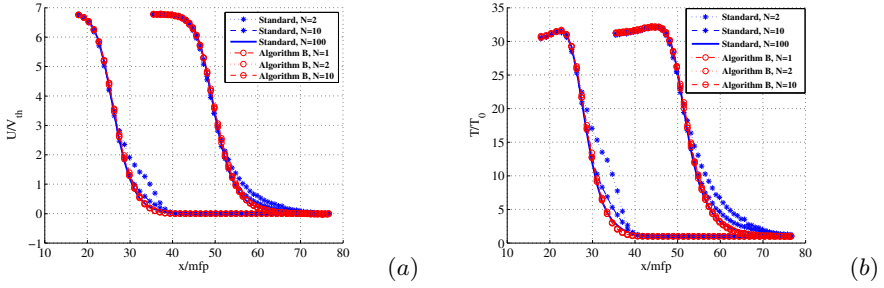


Fig. 1. Formation of a strong shock wave in front of a piston moving with a constant velocity of 2285.5 m/s in the x-direction; (a) comparison of velocity profiles at times $t = 1.0 \times 10^{-4}$ s and $t = 2.0 \times 10^{-4}$ s; (b) comparison of temperature profiles at the same times

collision scheme with $\langle N \rangle = 100.0$ particles per cell. The profiles computed by the NTC scheme with $\langle N \rangle = 10.0$ and $\langle N \rangle = 2.0$ (shown in stars) increasingly deviate from the reference profile in the lower part of the shock wave front, where the average number of particles per cell is smaller. The characteristic swelling in the lower part of the shock front is an effect of the insufficient number of effective collisions owing to an increase of repeated collisions of particle pairs with high relative velocity. Unlike these results, all profiles, obtained by algorithm B (shown in circles), are in a good agreement with the reference profiles. Figure 2

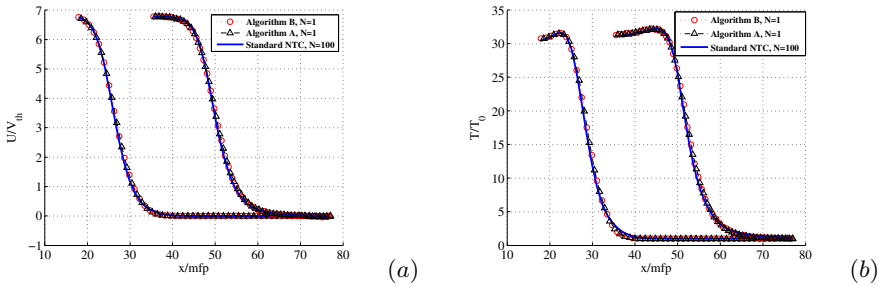


Fig. 2. Comparison of velocity (a) and temperature (b) profiles obtained from the calculation by using algorithm A and Algorithm B with a mean number of particles $\langle N \rangle = 1.0$ in cell

illustrates the comparison between the results obtained by algorithms A and B. The velocity and temperature profiles of both algorithms, A (triangles) and B (circles), obtained with very small mean number of particles per cell $\langle N \rangle = 1.0$, are in an excellent agreement with the reference NTC profiles, obtained with $\langle N \rangle = 100.0$ particles per cell.

4 Concluding Remarks

The comparison analysis on the shock wave formation problem demonstrates that, under certain conditions, the modified DSMC scheme can be used successfully for simulation of gas flows with very small numbers of particles per cell compared with the standard DSMC method. The simplified BT algorithm B has an efficiency of the same order such as the standard NTC collision algorithm.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement ITN GASMEMS n 215504. The author acknowledges the financial support provided by the NSF of Bulgaria under Grant No DID 02/20-2009.

References

1. Bird, G.A.: *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Clarendon Press, Oxford (1994)
2. Bird, G.A.: *Molecular Gas Dynamics*. Oxford University Press, Oxford (1976)
3. Koura, K.: Null-collision technique in the Direct Simulation Monte Carlo technique. *Phys. Fluids* 29, 3509–3511 (1986)
4. Yanitskiy, V.: Operator approach to Direct Simulation Monte Carlo theory in rarefied gas dynamics. In: Beylich, A. (ed.) *Proc. 17th Symp. on Rarefied Gas Dynamics*, pp. 770–777. VCH, New York (1990)
5. Babovsky, H.: On a simulation scheme for the Boltzmann equation. *Math. Methods Appl. Sci.*, 8, 223–233 (1986)
6. Ivanov, M., Rogasinsky, S.: Theoretical analysis of traditional and modern schemes of the DSMC method. In: Beylich, A. (ed.) *Proc. 17th Symp. on Rarefied Gas Dynamics*, pp. 629–642. VCH, New York (1990)
7. Dimov, I.: *Monte Carlo methods for applied scientists*. World Scientific, London (2008)
8. Strang, G.: On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.* 5, 506–517 (1968)
9. Stefanov, S., Roussinov, V., Cercignani, C.: Rayleigh-Bénard Flow of a Rarefied Gas and its Attractors. III. Three-dimensional Computer Simulations. *Phys. Fluids* 19, 124101 (2007)
10. Cercignani, C.: *The Boltzmann Equation and its Applications*. Springer, New York (1988)
11. Bobylev, A., Ohwada, T.: The error of the splitting scheme for solving evolutionary equations. *Appl. Math. Lett.* 14, 45–48 (2001)
12. Wagner, W.: A Convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation. *J. Stat. Phys.* 66, 1011–1044 (1992)
13. Kac, M.: *Probability and related topics in physical sciences*. Interscience Publishers Ltd., London (1959)

Is Self-Heating Important in Nanowire FETs?

D. Vasileska¹, A. Hossain², K. Raleva³, and S.M. Goodnick¹

¹ Arizona State University, Tempe, AZ 85287-5706, USA

² Intel Corp., Chandler, AZ, USA

³ University Sts, Cyril and Methodi, Skopje, Republic of Macedonia
vasileska@asu.edu

Abstract. In this work we investigate self-heating effects in nanowire FETs. We find that, as in the case of FD SOI devices, the velocity overshoot effect of the carriers in the channel and reduced number of scattering events with phonons lead to smaller ON-current degradation in short compared to long channel nanowire transistors.

Keywords: nanowire FETs, self-heating effects, particle-based device simulations, energy balance model for optical and acoustic phonons.

1 Introduction

In September of 2009, Intel illustrated the first working example of chips built on a 22 nanometer (nm) process [1]. A single example chip about the size of a fingernail contains about 2.9 billion transistors and about 364 megabits (45.5MB) of static RAM. Full 22nm production is expected in year 2011. A natural question to ask at this point is: How far down we can go with transistor scaling? Is 10nm technology going to make it on the market soon?

Because it is highly probable that conventional transistors can not be scaled anymore, researchers since more than a decade ago have started working on improving the performance of existing technology node in two ways. Approach 1 consists of utilization of alternative materials (gate stacks with high-k dielectrics which were already introduced in the 45nm technology node), strain and different silicon device crystallographic orientations. Uniaxial tensile strain has already shown benefit and larger speed for n-channel devices and biaxial compressive strain and SiGe channel have shown larger on-current for p-channel devices. However, this approach can make a push ahead of one or two technology nodes, not more. Because of that, in recent years researchers have worked extensively on Approach 2 which deals with alternative device designs, ranging from Fully-Depleted (FD) SOI devices, Dual-Gate (DG) structures, FinFETs or tri-gate structures, etc. All these devices that belong to the alternative device designs benefit from the (1) reduced junction capacitance, (b) absence of latch-up, (c) ease in scaling (buried oxide need not be scaled), (d) compatible with conventional silicon processing, (e) sometimes requires fewer steps to fabricate, (f) reduced leakage and (g) improvement in the soft error rate. Yet these structures as well face the problem of scaling. Possible alternatives as seen now are carbon nanotube, graphene and nanowire transistors.

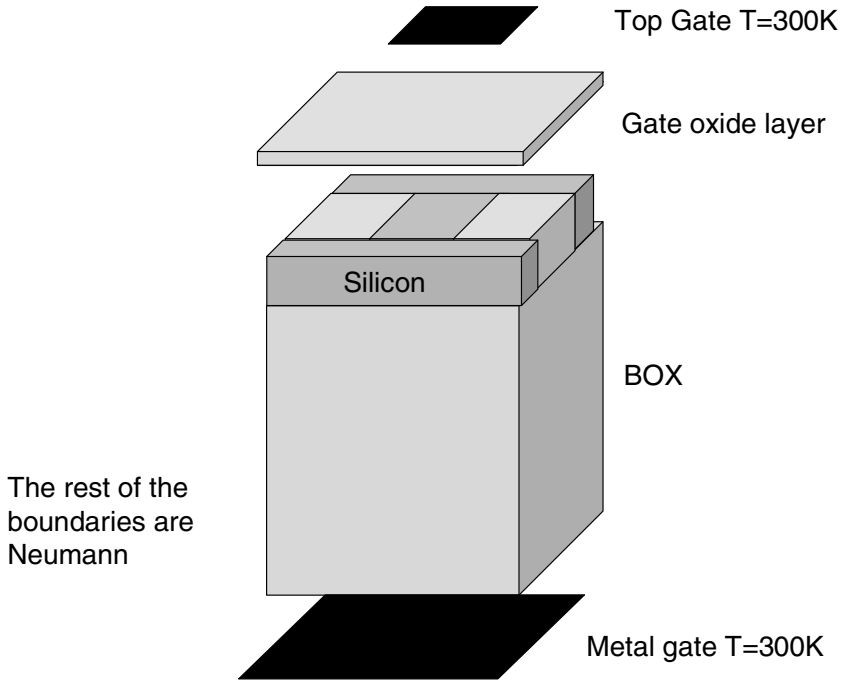


Fig. 1. Schematic description of the nanowire FET being simulated with our electro-thermal simulator

The nanowire transistors have many possible applications including being simply a FET or they can serve as biological biosensors. In the case when they are used as FETs large currents flow through the nanowire leading to self-heating effects. The self-heating effects in nanowire transistors arise from the fact that (1) the Buried Oxide Layer (BOX) serves as a thermal resistor to heat flow in the substrate contact (which is typically assumed at 300K), and also (2) the thermal conductivity of nanowires is much smaller than the thermal conductivity of bulk silicon or thin films that we encounter in FD SOI devices, due to the significant boundary scattering. Extensive experimental thermal conductivity measurements of silicon nanowires have been performed by Li Shi and co-workers [2] and thorough modeling efforts that explain the measured data of Li Shi et al. have been done by Mingo [3].

The purpose of this work is to examine whether self-heating effects will degrade the nanowire FET output characteristics (ON state) where they are expected to be most significant. To achieve this goal, our existing 2D electro-thermal device simulator [4,5,6] has been extended to three spatial dimensions and has the capability to study arbitrary material structures because of the positional dependence of the dielectric constant and of the thermal conductivity. This is the first device simulator that self-consistently solves the Boltzmann transport equation for the electrons (coupled to a 3D Poisson equation solver)

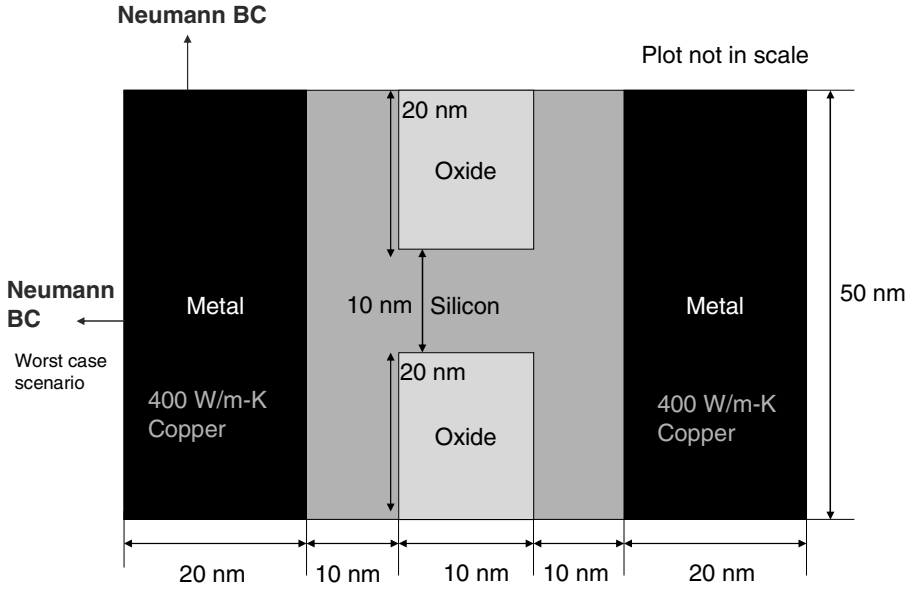


Fig. 2. Position of the metal gates and thermal boundary conditions in the device plane underneath the gate oxide

with a 3D energy balance solvers for the acoustic and the optical bath. It is also a difficult problem because one is coupling particle description for the electrons (which is inherently noisy) with a fluid description of the separate acoustic and optical phonon baths. Significant averaging and smoothing is necessary when passing the variables from the Monte Carlo to the energy balance solvers. Calculation details of the scheme developed at Arizona State University to couple the electron and phonon solvers can be found in Ref. [4,5,6]. In this paper we focus on the magnitude of the self-heating effects for different biasing conditions and on identifying what are the bottlenecks to heat flow in a particular nanowire structure. Details of the structure being examined, convergence plots, current degradation due to self-heating and lattice temperature profiles under different biasing conditions are presented in Section 2 of this paper. Conclusions regarding this work and future directions of research are given in Section 3.

2 Nanowire FET Electro-Thermal Simulations

The nanowire FET being simulated in this work is schematically illustrated in Fig. 1. The gate oxide is 0.8 nm thick and the BOX is 10 nm thick. The dimensions of the silicon nanowire are: 10 nm length, 7 nm thickness and 10 nm width. For the thermal conductivity that appears in the acoustic phonons energy balance solvers we have taken the value from Li Shi measurements [2] that correspond to wire with cross-section of 7-10 nm.

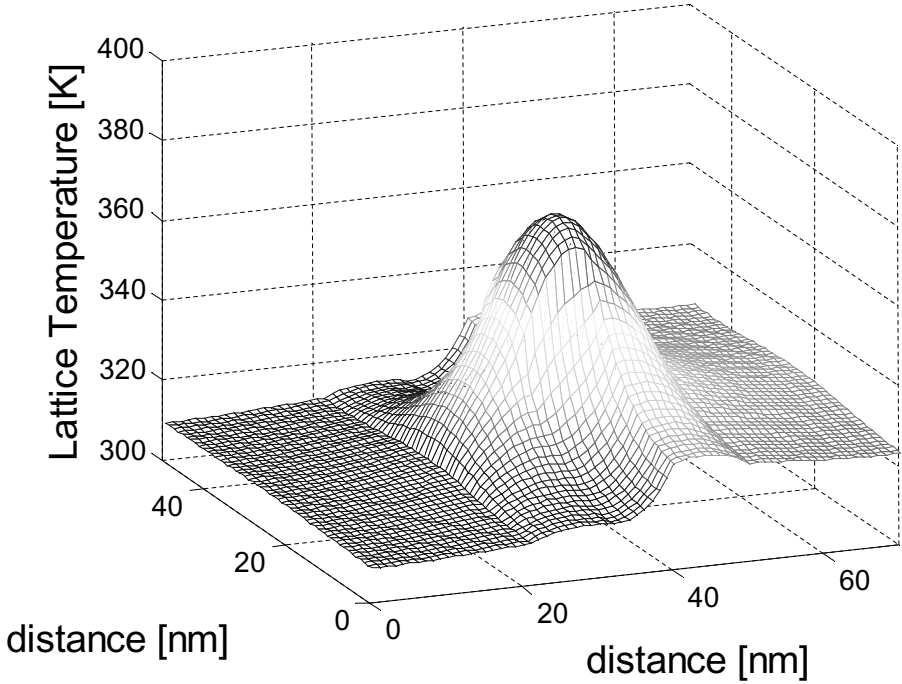


Fig. 3. Lattice temperature profile for $V_G = V_D = 1$ V

When solving the energy balance equations for the acoustic and optical phonons, boundary conditions on the lattice temperature must be established. Recalling that there is an analogy between electrical and thermal variables, from Ohm's law for the electrical conduction and Fourier law for heat conduction one immediately sees that electrostatic potential is analogous to lattice temperature and electrical current is analogous to heat flux. We know that when solving the Poisson equation for the electrostatic potential one has to define at least one node on the Poisson mesh with Dirichlet boundary conditions to connect to the outside world. Hence, in the lattice temperature mesh at least one node has to have Dirichlet boundary conditions. In all the simulations presented in this work, the bottom (substrate) electrode is taken to be at lattice temperature $T=300$ K and the top gate is also assumed to be at temperature $T=300$ K (see Fig. 1). For all the other boundaries, Neumann boundary conditions are assumed.

At the device plane under the gate oxide, we have assumed that we have copper metal gates dimensions of which are shown in Fig. 2. Copper has thermal conductivity of 400 W/m-K. Neumann boundary conditions are assumed at the edges of the metal boundaries.

The set of simulation results obtained with the electro-thermal particle-based device simulator are shown in Figs. 3, 4 and 5. In Fig. 3 we present the lattice temperature profile for bias conditions of $V_D = V_G = 1$ V. The optical

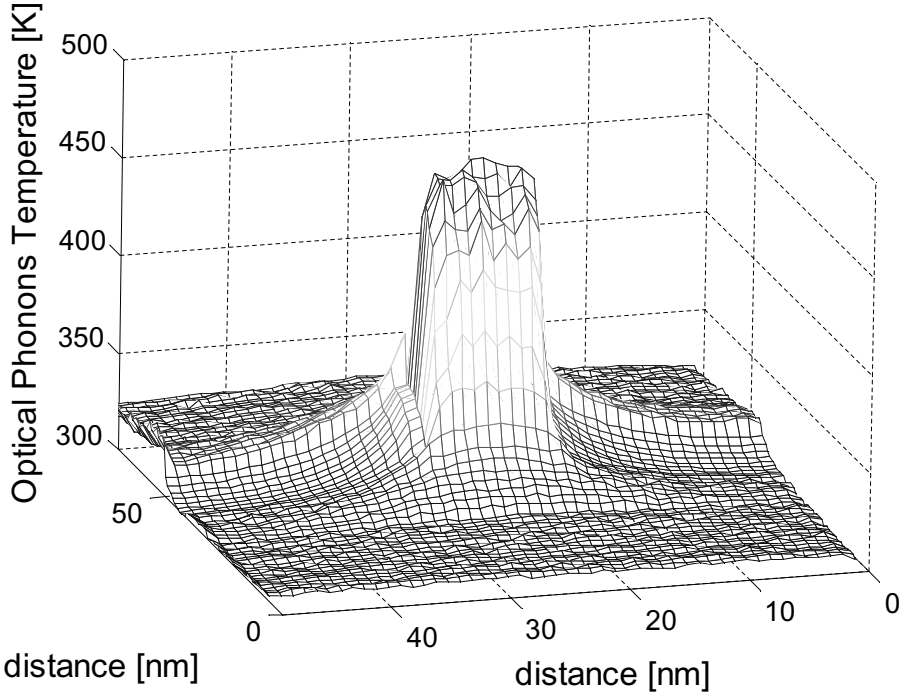


Fig. 4. Optical phonons profile for $V_G = V_D = 1$ V

phonon temperature for the same bias conditions is shown in Fig. 4. We see that at the drain end of the channel, where the electrons have the largest velocity and energy, there is a hot spot with peak optical phonon temperature of around 450 K. The acoustic phonon temperature is more smeared out and peak acoustic phonon temperature is smaller than the optical phonon temperature as it should be expected from a physical standpoint that the bottleneck in the heat transfer process are the optical, not the acoustic phonons because they have almost zero group velocity. The group velocity of the acoustic phonons equals the velocity of sound. The degradation of the current for these bias conditions is about 3.5% which is to be expected from the almost negligible degradation of the drift velocity of the carriers in the nanowire due to self-heating effects. The drift velocity data for the same bias conditions are shown in Fig. 5. We also want to point out that convergence in the current up to the third digit is achieved in 5 Gummel cycles.

3 Conclusions and Future Work

In summary, in this work we have presented preliminary simulation results for self-heating effects in nanowire transistors. As in the case of FD SOI devices, here as well, velocity overshoot leads to insignificant current degradation because

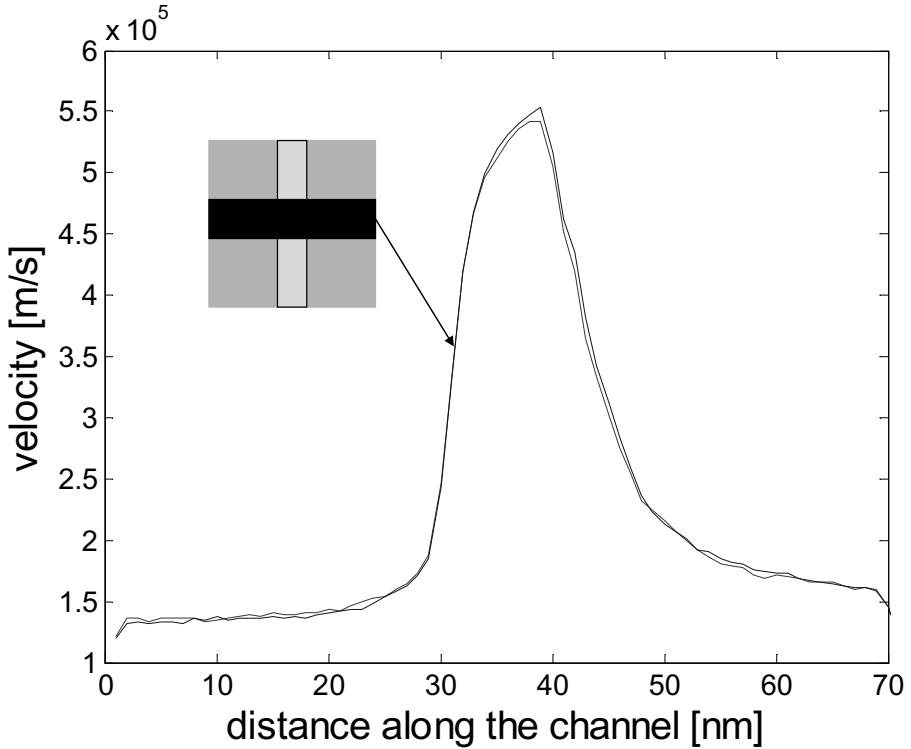


Fig. 5. Drift velocity profile for $V_G = V_D = 1$ V. This is the average velocity in the slab area denoted in the inset of the figure. The curve with maximum drift velocity corresponds to the isothermal case and the one with degraded drift velocity to the case when self-heating effects are included in the model.

of self-heating effects. There are many more issues that need to be addressed to derive even more conclusive results regarding self-heating in these nanowire transistors such as the inclusion of the temperature and position dependence of the thermal conductivity, to investigate the role of phonon boundary scattering in short wires and to examine whether phonon boundary scattering is sufficiently large in short nanowires so that the thermal conductivity itself has meaning. The later will require coupling of a phonon Boltzmann solver to the electron Boltzmann solver and that development work is currently underway at Arizona State University.

References

1. <http://www.intel.com>
2. Li, D., Wu, Y., Kim, P., Shi, L., Yang, P., Majumdar, A.: Thermal conductivity of individual silicon nanowires. *Appl. Phys. Lett.* 83, 2934–2936 (2003)

3. Mingo, N.: Calculation of Si nanowire thermal conductivity using complete dispersion relation. *Physical Review, B* 68, 113–308 (2003)
4. Raleva, K., Vasileska, D., Goodnick, S.M., Nedjalkov, M.: Modeling Thermal Effects in Nanodevices. *IEEE Trans. on Elec. Devices* 55(6), 1306 (2008)
5. Vasileska, D., Raleva, K., Goodnick, S.M.: Self-Heating Effects in Nano-Scale FD SOI Devices: The Role of the Substrate, Boundary Conditions at Various Interfaces and the Dielectric Material Type for the BOX. *IEEE Trans. Electron Devices* 56(12), 3064–3071 (2009)
6. Vasileska, D., Raleva, K., Goodnick, S.M.: Electrothermal Studies of FD SOI Devices That Utilize a New Theoretical Model for the Temperature and Thickness Dependence of the Thermal Conductivity. *IEEE Transactions on Electron Devices* 57, 726–728 (2010)

Mixed-Hybrid Formulation of Multidimensional Fracture Flow

Jan Březina and Milan Hokr

Technical University in Liberec, Studentská 2,
461 17 Liberec 1, Czech Republic
`jan.brezina@tul.cz`, `milan.hokr@tul.cz`

Abstract. We shall study Darcy flow on the heterogeneous system of 3D, 2D, and 1D domains and we present four models for coupling of the flow. For one of these models, we describe in detail its mixed-hybrid formulation. Finally, we show that Schur complements are suitable for solution of the linear system resulting from the lowest order approximation of the mixed-hybrid formulation.

Keywords: fracture flow, multidimensional coupling, Schur complement.

1 Introduction

The granite rock represents one of the suitable sites for a nuclear waste deposit. Water in the granite massive is conducted by the complex system of fractures of various sizes. While the small fractures can be modeled by an equivalent permeable continuum, the preferential flow in the large geological dislocations and their intersections should be considered as a 2D flow and 1D flow respectively. Motivated by this application, we have developed a simulator (Flow123d) of fracture flow and transport with a multidimensional coupling. Even after several successful applications of this model (e.g. [5]), there is a gap in its theoretical description. The aim of this work is to fill in this gap at least concerning the water flow.

In the second section, we shall present several conceptual models for the coupling between Darcian flows in different dimensions. Then we select one of these models and set up a fully coupled 1-2-3 dimensional problem. In the third section we describe the mixed-hybrid (MH) formulation of the fully coupled problem. We basically follow Maryška, Rozložník, Tůma [6] and Arbogast, Wheeler, Zhang [1], but we rather derive MH-formulation as an abstract saddle point problem in order to use classical theory due to Brezzi and Fortin [2]. Finally, in Section 4 we use Schur complements to solve the linear system resulting from the discretization. We shall prove key properties of the Schur complements similarly as in [7] and we confirm these properties by numerical experiments.

2 Physical Setting

Common model of the underground water flow is the continuity equation

$$\operatorname{div} \mathbf{v} = f \quad (1)$$

completed by Darcy's law

$$\mathbf{v} = -\mathbb{K} \nabla h, \quad (2)$$

where $\mathbf{v}$ is the Darcy flux [ms^{-1}], h is the water pressure head [m], f is the volume density of the water sources [s^{-1}], and $\mathbb{K}$ is the tensor of hydraulic conductivity [ms^{-1}]. Let us consider the water flow described by (1), (2) in a 3D porous medium that contains very thin layers and channels with a substantially different hydraulic conductivity. Due to the different conductivity these features can not be neglected, but can be considered as 2D and 1D objects respectively. We denote $\Omega_3 \subset \mathbf{R}^3$ the 3D domain, $\Omega_2 \subset \Omega_3$ will be the domain of 2D fractures, and $\Omega_1 \subset \Omega_2$ is the domain of 1D channels. In order to keep further formulas consistent, we also introduce Ω_0 as the set of channel intersections. Since the fractures and channels are thin, we can assume that the velocity and the pressure is constant in the normal direction. Moreover the normal part of the velocity can be interpreted as the water interchange with the surrounding medium. Consequently we can integrate (1) along the normal directions and obtain

$$\operatorname{div} \mathbf{q}_d = F_d \quad \text{on } \Omega_d \setminus \Omega_{d-1} \quad \text{for } d \in \{1, 2, 3\}, \quad (3)$$

where $\mathbf{q}_3 = \mathbf{v}_3$ is simply the Darcy flux [ms^{-1}], $\mathbf{q}_2 = \delta_2 \mathbf{v}_2$ [$m^2 s^{-1}$] is the water flux through the 2D fracture of thickness δ_2 [m], and $\mathbf{q}_1 = \delta_1 \mathbf{v}_1$ [$m^3 s^{-1}$] is the water flux through the 1D channel of cross-section δ_1 [m^2]. Further, F_d are partially integrated densities of the water sources, which we shall discuss presently. Vectors $\mathbf{q}_d$ and tensors $\mathbb{K}_d$, $d \in \{1, 2, 3\}$ lives in the corresponding tangent space of Ω_d . Similarly, we denote h_d the pressure head on the domain Ω_d .

Next, we have to introduce suitable coupling between the equations on the domains of different dimension. We assume that the water flux q_{ab} from Ω_a to Ω_b is driven by the pressure head difference:

$$q_{ab} = \sigma_{ab}(h_a - h_b), \quad (4)$$

where σ_{ab} is an water transition coefficient. However, there are at least four different models for 2D-1D and 3D-2D interaction based on the equation (4).

Let us explain it on 2D-1D case (see Figure 2). We can choose either discontinuous or continuous pressure head on 2D. In the first case there is one independent water interchange for each of two sides of the 1D domain and the 2D pressure head is discontinuous over the 1D fracture. That is why we call it also a *separating fracture*. In the second case we assume continuous 2D pressure and only one total flux between 2D and 1D.

Independently, we can choose either communication over the volume or over the surface. In the case of the volume communication the flux q_{ab} acts as a volume source [s^{-1}] in both dimension. Nevertheless, to keep it constant in the normal

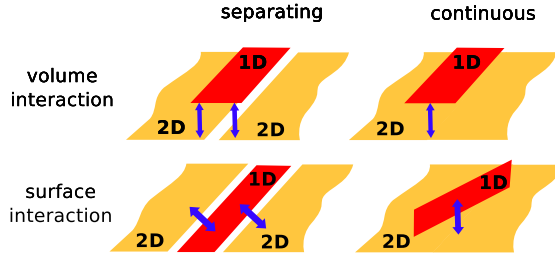


Fig. 1. Four possible interaction models between 2D and 1D

direction of the 1D domain, we have to perform averaging of q_{ab} over the width δ [m] of the 1D domain. The transition coefficient σ has unit $[m^{-1}s^{-1}]$ and has the same meaning as the water transfer coefficient in the dual continuum models (see [3]). In the case of the surface communication, the outflow q_{ab} [ms^{-1}] from the boundary of the 2D domain spreads over the width δ [m] of the 1D domain so that q_{ab}/δ act as a volume source in the 1D domain. The transition coefficient σ [s^{-1}] should be proportional to $|\mathbb{K}|\delta$.

In what follows, we consider only the model with discontinuous pressure head and the surface communication. On the domain Ω_2 , there is one water outflow from Ω_3 for every side of the surface:

$$\begin{aligned} \mathbf{q}_3 \cdot \mathbf{n}^+ &= q_{32}^+ = \sigma_3^+(h_3^+ - h_2), \\ \mathbf{q}_3 \cdot \mathbf{n}^- &= q_{32}^- = \sigma_3^-(h_3^- - h_2), \end{aligned}$$

where $\mathbf{q}_3 \cdot \mathbf{n}^{+/-}$ [ms^{-1}] is the outflow from Ω_3 , $h_3^{+/-}$ [m] is the trace of the pressure head on Ω_3 , h_2 [m] is the pressure head on Ω_2 , and $\sigma_3^{+/-} = \sigma_{32}$ [s^{-1}] is the transition coefficient. On the other hand, the sum of the interchange fluxes $q_{32}^{+/-}$ forms a volume source on Ω_2 . Therefore F_2 [ms^{-1}] on the right hand side of (3) is given by

$$F_2 = \delta_2 f_2 + (q_{32}^+ + q_{32}^-). \quad (5)$$

The communication between Ω_2 and Ω_1 is similar. However, in the 3D ambient space, an 1D channel can adjoin multiple 2D fractures $1, \dots, n$. Therefore, we have n independent outflows from Ω_2 :

$$\mathbf{q}_2 \cdot \mathbf{n}^i = q_{21}^i = \sigma_2^i(h_2^i - h_1),$$

where $\sigma_2^i = \delta_2^i \sigma_{21}$ [ms^{-1}] is the transition coefficient integrated over the width of the fracture i . Sum of the fluxes forms F_1 [$m^2 s^{-1}$]

$$F_1 = \delta_1 f_1 + \sum_i q_{21}^i. \quad (6)$$

For the consistency we also set $F_3 = f_3$ [s^{-1}], $\delta_3 = 1$ [-], and $\sigma_1 = 0$.

In order to obtain unique solution we have to prescribe boundary conditions. We assume that $\partial\Omega_1 \subset \partial\Omega_2 \subset \partial\Omega_3$. Let us denote Γ_d^D the Dirichlet part of the boundary $\partial\Omega_d$, where we prescribe the pressure head P_d . On the remaining part Γ_d^W , we prescribe outflow by the Newton boundary condition

$$\mathbf{q}_d \cdot \mathbf{n} = \alpha_d(h_d - P_d^W).$$

where $\alpha_3 [s^{-1}]$, $\alpha_2 [ms^{-1}]$, $\alpha_1 [m^2s^{-1}]$ are a transition coefficients and P_d^W is the given outer pressure head.

3 Mixed-Hybrid Formulation of Multidimensional Fracture Flow Problem

Now, we are going to introduce MH-formulation of the problem denoted in the previous section. To avoid technicalities, we assume that Ω_3 have piece-wise polygonal boundary, domain Ω_2 consists of polygons, and Ω_1 consists of line segments. We also assume $\partial\Omega_1 \subset \partial\Omega_2 \subset \partial\Omega_3$. Further, we decompose Ω_d , $d \in \{1, 2, 3\}$ into sub-domains Ω_d^i , $i \in I_d$ satisfying the compatibility condition

$$\Omega_{d-1} \subset \Gamma_d \setminus \partial\Omega_d, \quad d = 1, 2, 3 \quad \text{where } \Gamma_d = \bigcup_{i \in I_d} \partial\Omega_d^i. \quad (7)$$

The idea of MH-formulation is to integrate (2) by parts on every sub-domain. There appears a term with the trace of the pressure head, which is considered as a Lagrange multiplier to enforce continuity of the pressure head over the boundaries. However, since the pressure head could be discontinuous over the fractures, we have to deal with two distinct multipliers along Ω_2 and Ω_1 . To this end, we introduce a natural decomposition Ω_d^j , $j \in J_d$ with boundaries given by Ω_{d-1} . Due to the compatibility condition (7) the decomposition I_d can be viewed as a refinement of the decomposition J_d . In particular, for every Ω_d^i , $i \in I_d$ there is a unique $j(i)$ such that $\Omega_d^i \subset \Omega_d^{j(i)}$. Then the Lagrange multiplier for the sub-domain Ω_d^j , $j \in J_d$ have support on the set

$$\Gamma_d^j = \Gamma_d \cap \overline{\Omega_d^j}. \quad (8)$$

Following [1] and [6], we shall consider following spaces for the MH-solution:

$$V = V_3 \times V_2 \times V_1 = \prod_{d \in \{3, 2, 1\}} \prod_{i \in I_d} H(\text{div}, \Omega_d^i), \quad (9)$$

$$P = P_3 \times P_2 \times P_1 \times \mathring{P}_3 \times \mathring{P}_2 \times \mathring{P}_1, \quad (10)$$

$$P_d = L^2(\Omega_d), \quad \mathring{P}_d = \prod_{j \in J_d} \left\{ \mathring{\varphi} \in H^{1/2}(\Gamma_d^j) \mid \mathring{\varphi} = 0 \text{ on } \Gamma_d^D \right\}.$$

where $H(\text{div}, \Omega)$ is standard space of L^2 -vector functions with divergence in $L^2(\Omega)$, and $H^{1/2}(\partial\Omega)$ is the space of traces of functions from $H^1(\Omega)$. In the

definition of the MH-solution, the flux $\mathbf{q}_d$ is from V_d , the pressure head h_d from P_d and the Lagrange multiplier or the pressure head trace $\tilde{h}$ is from $\tilde{P}_d$. Introduction of the composed spaces V and P allows us to formulate MH-problem as an abstract saddle problem in the spirit of [2].

Definition 1. We say that pair $(\mathbf{q}, h) \in V \times P$ is MH-solution of the problem if it satisfy abstract saddle point problem

$$a(\mathbf{q}, \psi) + b(\psi, h) = \langle F, \psi \rangle \quad \forall \psi \in V, \quad (11)$$

$$b(\mathbf{q}, \varphi) - c(h, \varphi) = \langle G, \varphi \rangle \quad \forall \varphi \in P, \quad (12)$$

where bilinear forms on the left-hand side are

$$\begin{aligned} a(\mathbf{q}, \psi) &= \sum_{d=1,2,3} \sum_{i \in I_d} \int_{\Omega_d^i} \frac{1}{\delta_d} \mathbf{q}_d^i \mathbb{K}_d^{-1} \psi_d^i, \\ b(\mathbf{q}, \varphi) &= \sum_{d=1,2,3} \sum_{i \in I_d} \left(\int_{\Omega_d^i} -\operatorname{div} \mathbf{q}_d^i \varphi_d + \int_{\partial \Omega_d^i} (\mathbf{q}_d^i \cdot \mathbf{n}) \tilde{\varphi}_d^{j(i)} \right), \\ c(h, \varphi) &= \sum_{d=1,2,3} \sum_{j \in J_d} \left(\int_{\Gamma_d^j \cap \Omega_{d-1}} \sigma_d (h_{d-1} - \tilde{h}_d^j) (\varphi_{d-1} - \tilde{\varphi}_d^j) + \int_{\Gamma_d^j \cap \Gamma_d^W} \alpha_d \tilde{h}_d^j \tilde{\varphi}_d^j \right), \end{aligned}$$

and linear forms on the right-hand side are

$$\begin{aligned} \langle G, \psi \rangle &= \sum_{d=1,2,3} \sum_{i \in I_d} \int_{\partial \Omega_d^i} \tilde{P}_d(\psi_d \cdot \mathbf{n}), \\ \langle F, \varphi \rangle &= - \sum_{d=1,2,3} \left(\int_{\Omega_d} \delta_d f_d \varphi_d + \sum_{j \in J_d} \int_{\Gamma_d^j \cap \Gamma_d^W} \alpha_d P_d^W \tilde{\varphi}_d^j \right). \end{aligned}$$

where $\tilde{P}_d \in \tilde{P}_d$ is any extension of the Dirichlet condition $P_d \in H^{1/2}(\Gamma_d^D)$. Consequently the full trace of the unknown pressure head is $\tilde{h}_d + \tilde{P}_d$.

The second term of the form b deserves a note. The outflow $\mathbf{q}_d^i \cdot \mathbf{n}$ is from dual to $H^{1/2}(\partial \Omega_d^i)$ which in general is not subspace of $H^{1/2}$ on the larger domain, namely $\Gamma_d^{j(i)}$. But here we use the fact, that the later domain does not penetrate into the domain Ω_d^i .

Assuming that δ_d , $\mathbb{K}_d$, σ_d , and α_d are uniformly bounded and uniformly grater than zero (positive definiteness of $\mathbb{K}_d$), we can prove that $a(\cdot, \cdot)$ and $c(\cdot, \cdot)$ are bounded, symmetric, positive definite bilinear forms and that

$$\mathcal{B} : V \rightarrow P', \quad \langle \mathcal{B}(\mathbf{q}, \varphi) \rangle = b(\mathbf{q}, \varphi)$$

is surjective operator. Assuming further

$$f_d \in L^2(\Omega_d), \quad P_d \in H^{1/2}(\Gamma_d^D), \quad P_d^W \in L^2(\Gamma_d^W),$$

we can prove that the MH-solution is independent of choice of decomposition I_d and independent of choice of extension $\tilde{P}_d$. Finally, using [2, Theorem 1.2], we can prove existence and uniqueness of the MH-solution.

4 Linear System and Its Schur Complements

The advantage of a discretization based on the mixed-hybrid formulation is a particular form of the resulting linear system, which could be effectively solved by Schur complements. This shall be investigated in this section.

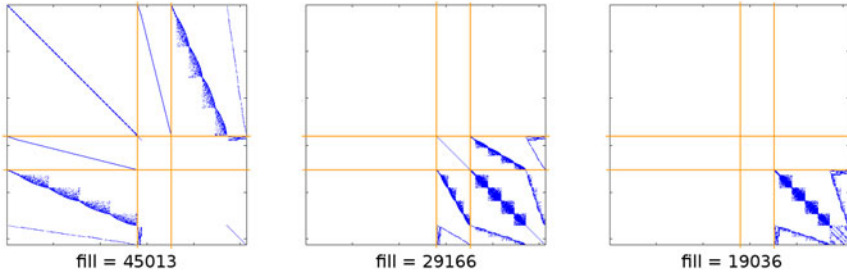


Fig. 2. Sparsity pattern: MH matrix (a), the first Schur complement (b), and the second (c). “fill” denotes number of non-zero entries, see also Table 1.

We consider the lowest order approximation of the MH-formulation. To this end, we choose simplicial elements as the sub-domains Ω_d^i , $i \in I_d$. Then, we approximate the space $H(\text{div}, \Omega_d^i)$ by the Raviart-Thomas space $RT_0(\Omega_d^i)$ (see [2]) and the spaces $L^2(\Omega_d)$ and $H^{1/2}(\Gamma_d^j)$ by piecewise constant functions on elements and their edges respectively (for details see [6]). Such discretization leads to the linear system which inherits the saddle-point structure of the equations (11), (12). The system matrix $\mathbb{A}$ has a block structure (see also Figure 2).

$$\mathbb{A} = \begin{pmatrix} A & B^T & \mathring{B}^T \\ B & C & \mathring{C}^T \\ \mathring{B} & \mathring{C} & \tilde{C} \end{pmatrix}$$

Full analysis of the system matrix and its Schur complements for a 3D domain and prismatic finite elements was done by MARYŠKA, ROZLOŽNÍK, and TŮMA in [7]. For our multidimensional case, we only mention the main properties. Block A is discrete version of $a(\cdot, \cdot)$ and consists of positive-definite blocks $(d+1) \times (d+1)$ on the diagonal for every d -dimensional mesh element. Therefore, the inverse A^{-1} is also positive-definite and easy to compute. Consequently, we can form the first Schur complement

$$\mathbb{A}_1 = \mathbb{A}/A = C - (B, \mathring{B})^T A^{-1} (B, \mathring{B})$$

in the linear time with respect to the problem size. Moreover, the blocks $C, \mathring{C}, \tilde{C}$ are discretizations of the form $-c(\cdot, \cdot)$, thus whole C -block is negative-definite and $\mathbb{A}/A$ is positive-definite.

The block B come from the first term of the form $b(\cdot, \cdot)$. For every d -dimensional element there is a row with $(d+1)$ non-zeroes located in the columns of the corresponding diagonal block in the matrix A . Using this property, we observe that leading block of $\mathbb{A}_1$, i.e. $A_1 = B^T A^{-1} B$, is diagonal. Hence we can compute the second Schur complement $\mathbb{A}_2 = (-\mathbb{A}_1)/A_1$ in the linear time as well. Further, we shall prove that $\mathbb{A}_2$ is positive-definite by showing that the Schur complement of any positive definite matrix is also positive definite. Let

$$M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$$

be positive definite. One can check that

$$M^{-1} = \begin{pmatrix} A^{-1} + (A^{-1}B)\overline{C}(A^{-1}B)^T, & -(A^{-1}B\overline{C}) \\ -(A^{-1}B\overline{C})^T, & \overline{C} \end{pmatrix}, \quad \overline{C} = (M/A)^{-1}.$$

In particular $(M/A)^{-1}$ is principal sub-matrix of M^{-1} and thus have the interlacing property:

Proposition 1. [4, Theorem 8.1.7] *Let $B \in \mathbf{R}^{k \times k}$ be symmetric principal sub-matrix of a symmetric matrix $A \in \mathbf{R}^{n \times n}$. Denoting α_i and β_i decreasing eigenvalues sequence of A and B respectively, it holds*

$$\alpha_i \geq \beta_i \geq \alpha_{i+n-k}, \quad i = 1, \dots, k.$$

Consequently the smallest eigenvalue of M/A is bounded from below by the smallest eigenvalue of M .

Table 1. Size, the number of nonzero elements and condition number of the original MH matrix of problem P and its two Schur complements depicted on Figure 2

Schur complement	size	fill	condition number
$\mathbb{A}$	10258	45013	9.8e+05
$\mathbb{A}_1$	4662	29166	10.3e+05
$\mathbb{A}_2$	3218	19036	1.1e+05

We shell demonstrate our theoretical developments on a test problem P — a cube cut by two diagonal planes (fractures) into four prisms, Dirichlet boundary condition. In Figure 2 you can see block structure and sparsity pattern of the MH matrix $\mathbb{A}$ and Schur complements $\mathbb{A}_1, \mathbb{A}_2$ for a particular discretization using 1444 elements. Table 1 summarizes basic numerical properties of the matrices $\mathbb{A}, \mathbb{A}_1, \mathbb{A}_2$. In particular for $\mathbb{A}_2$, we get reduction of the size by factor of 3, reduction of the fill by factor of 2, and reduction of the condition number by factor of 10.

Better numerical properties of Schur complements should lead to better performance of the solver. We used BiCGStab method with stopping accuracy 10^{-7}

Table 2. Convergence comparison. Optimal factor level of ILU preconditioner using BiCGStab solver with accuracy 10^{-7} .

	112 755 elements			290 281 elements		
Schur complement	$\mathbb{A}$	$\mathbb{A}_1$	$\mathbb{A}_2$	$\mathbb{A}$	$\mathbb{A}_1$	$\mathbb{A}_2$
optimal factor level	9	3	2	13	3	3
iterations	45	31	44	42	46	49
solver time	40.4 s	18.6 s	15.4 s	118 s	72 s	63 s

preconditioned by ILU with factor level k to solve the systems $\mathbb{A}$, $\mathbb{A}_1$, $\mathbb{A}_2$ for two different discretizations of problem P. For every of these six systems, we found the optimal k that leads to the shortest solution time (higher values of k offer better preconditioner, but slower to apply because of higher fill of the matrix). Table 2 reports optimal values of k , number of iterations, and solution time including construction of Schur complements and preconditioner. The solution times for $\mathbb{A}_2$ are about half of the $\mathbb{A}$ case. Important observation is that in contrast to the matrix $\mathbb{A}$, the optimal factor level $k = 3$ for the Schur complements is independent of the problem size. Moreover, for $\mathbb{A}_1$, $\mathbb{A}_2$ one can use CG solver to get even shorter solution times.

Acknowledgment. This work is supported by the project 205/09/P657 of the Czech Science Foundation and by the research center ARTEC, project 1M0554 of MŠMT, Czech Republic.

References

1. Arbogast, T., Wheeler, M.F., Zhang, N.-Y.: A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM Journal on Numerical Analysis* 33(4), 1669–1687 (1996)

2. Fortin, M., Brezzi, F.: *Mixed and Hybrid Finite Element Methods*. Springer, Heidelberg (December 1991)

3. Gerke, H.H., van Genuchten, M.T.: Evaluation of a First-Order water transfer term for variably saturated Dual-Porosity flow models. *Water Resources Research* 29(4), 1225–1238

4. Golub, G.H., Van Loan, C.F.: *Matrix Computations* (Johns Hopkins Studies in Mathematical Sciences, 3rd edn. The Johns Hopkins University Press, Baltimore (October 1996)

5. Maryška, J., Severýn, O., Vohralík, M.: Numerical simulation of fracture flow with a mixed-hybrid FEM stochastic discrete fracture network model. *Computational Geosciences* 8(3), 217–234 (2005)

6. Maryška, J., Rozložník, M., Tůma, M.: Mixed-hybrid finite element approximation of the potential fluid flow problem. *J. Comp. Appl. Math.* 63, 383–392 (1995)

7. Maryška, J., Rozložník, M., Tůma, M.: Schur complement systems in the Mixed-Hybrid finite element approximation of the potential fluid flow problem. *SIAM J. Sci. Comput. (SISC)* 22(2), 704–723 (2000)

WRF-Fire Applied in Bulgaria

Nina Dobrinkova¹, Georgi Jordanov², and Jan Mandel³

¹ Institute of Information and Communication Technologies
Bulgarian Academy of Sciences

`nido@math.bas.bg`

² Geophysical Institute
Bulgarian Academy of Sciences

`gjordanov@geophys.bas.bg`

³ Department of Mathematical and Statistical Sciences
University of Colorado Denver

`jan.mandel@ucdenver.edu`

Abstract. WRF-Fire consists of the WRF (Weather Research and Forecasting Model) coupled with a fire spread model, based on the level-set method. We describe a preliminary application of WRF-Fire to a forest fire in Bulgaria, opportunities for research of forest fire models for Bulgaria, and plans for the development of an Environmental Decision Support Systems which includes computational modeling of fire behavior.

Keywords: Wildland fire modeling, forest fires, coupled atmosphere-fire modeling, level-set method, Decision Support System.

1 Introduction

Forest fires are a problem in most south-European countries, because of the dry climate during the summer and the year-round high temperatures. Statistics have been done among the south-European EU member states, where is clear that the number of the forest fires has increased rapidly in the last 15 years along with the climate change. In addition, increased pace of development puts more people and property into harm's way in a wildfire. Bulgaria as part of this region also has huge problems with wildland fires. Statistics have been maintained in Bulgaria for the last 30 years, and the number of forest fires increases every year [6,10]. Even though the number of wildfires is increasing and the consequences are not only of environmental, but also of economical and social significance, a proper solution has not been found neither for Bulgaria nor for the rest of the south EU member states. At present, most of the countries suffering from wildfires are dealing with the disaster at the moment of its occurrence. After much controversy, the role of forest fires as a natural part of the ecosystem and the importance of the fuel accumulation were recognized in North America, where prescribed burns are now an integral part of forest management to reduce the fuel, and software tools can play an important role in the fire management as well in the evaluation of prescribed burns [7]. Decision support tools integrating

models and observations from a variety of sources are of great interest in Bulgaria also [5].

Bulgaria has started a nice initiative by opening a center in Sofia on 7th of July 2007 to provide the disaster control units with adequate, real time information and to ensure better coordination and effectiveness in the prevention of natural hazards. It is the first initiative in the policy makers sector with such orientation. The center is named Aero-Spatial Observation Center (ASOC). It is aimed to improve and streamline the process of early warning, prediction and monitoring of natural disasters and accidents on a national scale. It allows discovering and following the dynamics of wildland and forest fires, floods, to estimate the loss of forest, control the conditions of the vegetation, soil humidity and erosion as well as air pollution. The system operates on national level and is also used by other governmental organizations e.g. Ministry of Economy and Energy, Ministry of Environment and Water Supplies, Ministry of Agriculture and Forestry, and others. The center operates 24/7. However ASOC still lacks a system for automated satellite image recognition and early detection of forest fires, floods, and other natural or human-caused disasters. That is why our team from the Bulgarian Academy of Sciences has started its own research project dedicated to the forest fires models and tools for fire simulations.

After an analysis of the literature, we have identified WRF-Fire [9] as a free Linux-based model which can simulate open area fires not only in the U.S., but also in Bulgaria, when the run process is modified to ingest data available in Europe. WRF contains the WRF Preprocessing System (WPS) [14, Chapter 3], which can input meteorological and land-use data in a number of commonly used formats. WPS has been extended to process fine-scale land data for use with the fire model, such as topography and fuel [2] [14, Appendix A]. While the format of meteorological data has largely stabilized, the ingestion of fire-modeling data was developed for U.S. sources only, and it may require further preprocessing for other countries.

2 Coupled Atmosphere-Fire Modeling by WRF-Fire

This section is based on [9], where more details can be found.

Fire models range from simple spread formulas to sophisticated computational fluid dynamics and combustion simulations, see the review in [13], and also [9, p. 50]. However, a fire behaviour model in a Decision Support System should be faster than real time in order to deliver a prediction, which dictates a compromise between the spatial resolution, the processes to be modeled, and fast execution.

Weather has a major influence on wildfire behavior; in particular, the wind plays a dominant role in the fire progress and shape. Conversely, the fire influences the weather through the heat and vapor fluxes. Fire heat output can easily reach the surface intensity of 1MW/m^2 , and the fast-rising hot air causes a significant air motion, which affects the atmosphere also away from the fire. It is known that a large fire “creates its own weather.” The correct wildland fire shape and progress result from the two-way interaction between the fire and the atmosphere [3,4].

2.1 Overview of the Software

WRF-Fire [9] combines the Weather Research and Forecasting Model (WRF) [15] with a semi-empirical fire spread model. WRF-Fire got its start in [12], where a combination of the tracer-based model from [3] with WRF was proposed, a road map was formulated, and the fundamental observation was made that the innermost domain, which interacts directly with the fire model, needs to run in the Large Eddy Simulation (LES) mode. However, instead of tracers, the fire code in WRF-Fire was developed [9] based on the level-set method [11], partly because the level-set function can be manipulated more easily than tracers for the purposes of data assimilation. The code in WRF-Fire for the fire spread rate and feedback to the atmosphere was taken from [3,4] without any significant changes, and the initial code for the WRF interface was taken from [12].

In the semi-empirical model, the fire spread rate in the normal direction to the fireline is assumed to be a function of the fuel properties, the wind speed close to the ground, and the terrain slope. The fraction of the fuel left is assumed to be an exponential function of the time from ignition. The semi-empirical formulas were derived from laboratory experiments, and the coupled model was verified on several large fires in an earlier implementation, called CAWFE [4], with the fire propagation by tracers and atmospheric modeling by the Clark-Hall weather code. WRF-Fire takes advantage of this validation and implements a subset of the physical model from [3,4]: the physical model is the same, but the fire spread in WRF-Fire is implemented by the level-set method, and the weather model is replaced by WRF, a supported standard community weather code. WRF can be run with several nested refined meshes, called domains in meteorology, which can run different physical models. WRF-Fire takes advantage of the mature WRF infrastructure for parallel computing and for data management. An important motivation for the development of the WRF-Fire software was the ability of WRF to export and import state, thus facilitating data assimilation (input of additional data while the model is running), which is essential for fire behaviour prediction from all available data [8].

2.2 Mathematical Methods

Mathematically, the fire model is posed in the horizontal (x, y) plane. The semi-empirical approach to fire propagation used here assumes that the fire spreads in the direction normal to the fireline at the speed given by the modified Rothermel's formula

$$S = \min\{B_0, R_0 + \phi_W + \phi_S\}, \quad (1)$$

where B_0 is the backing rate (spread rate against the wind), R_0 is the spread rate in the absence of wind, $\phi_W = a(\mathbf{v} \cdot \mathbf{n})^b$ is the wind correction, and $\phi_S = d\nabla z \cdot \mathbf{n}$ is the terrain correction. Here, $\mathbf{v}$ is the wind vector, ∇z is the terrain gradient vector, and $\mathbf{n}$ is the normal vector to the fireline in the direction away from the burning area. In addition, the spread rate is limited by $S \leq S_{\max}$. Once the fuel is ignited, the amount of the fuel at location (x, y) is given by

$$F(x, y, t) = F_0(x, y)e^{-(t-t_i(x, y))/T(x, y)}, \quad t > t_i(x, y) \quad (2)$$

where t is the time, t_i is the ignition time, F_0 is the initial amount of fuel, and T is the time constant of fuel (the time for the fuel to burn down to $1/e$ of the original quantity). The coefficients B_0 , R_0 , a , b , d , $S_{\max}$, F_0 , and T in (1) and (2) are data.

The heat fluxes from the fire are inserted into the atmospheric model as forcing terms in the differential equations of the atmospheric model into a layer above the surface, with exponential decay with altitude. The sensible heat flux is inserted as the time derivative of the temperature, while the latent heat flux as the time derivative of water vapor concentration. This scheme is required because atmospheric models with explicit timestepping, such as WRF, do not support flux boundary conditions. The heat fluxes from the fire to the atmosphere are taken proportional to the fuel burning rate, $\partial F(x, y, t)/\partial t$. The proportionality constants are again fuel coefficients.

For each point in the plane, the fuel coefficients are given by one of the 13 Anderson categories [1]. The categories are developed for the U.S. and different countries use different fuel schemes. WRF-Fire provides for the definition of the categories as input data, which allows the software to adapt to other countries.

The burning region at time t is represented level set function ϕ by a as the set of all points (x, y) where $\phi(x, y, t) < 0$. It is known that the level set function satisfies the partial differential equation [11]

$$\partial\phi/\partial t = -S|\nabla\phi|, \quad (3)$$

where $|\nabla\phi|$ is the Euclidean norm of the gradient of ϕ . Equation (3) is solved numerically by the finite difference method.

In each time step of the atmospheric model, first the winds are interpolated from the atmospheric model grid to a finer fire model grid. The numerical scheme for the level set equation (3) is then advanced to the next time step value, the time of ignition is set for any nodes that started burning during the time step, and the fuel burned during the time step is computed by quadrature from (2) in each fire model cell. The resulting heat fluxes are averaged over the fire cells that make up one atmosphere model cell, and inserted into the atmospheric model, which then completes its own time step.

3 Initialization and Computational Results

WRF-Fire v.3.2 is used for the simulation. The model consists of one domain of size 4 by 4 km, with horizontal resolution of 50 m for the atmosphere mesh, 80 by 80 grid cells, and with 41 vertical levels from ground surface to 100hPa. There is no nesting. The domain is located 4 km west from village Zheleznitsa in the south-east part of Sofia district. This domain is covering the low part of the forest on Vitosha mountain. The ignition line is located in the center of the domain, it is 345 m long and the ignition is made at 01 APR 2009, 06:00:02UTC

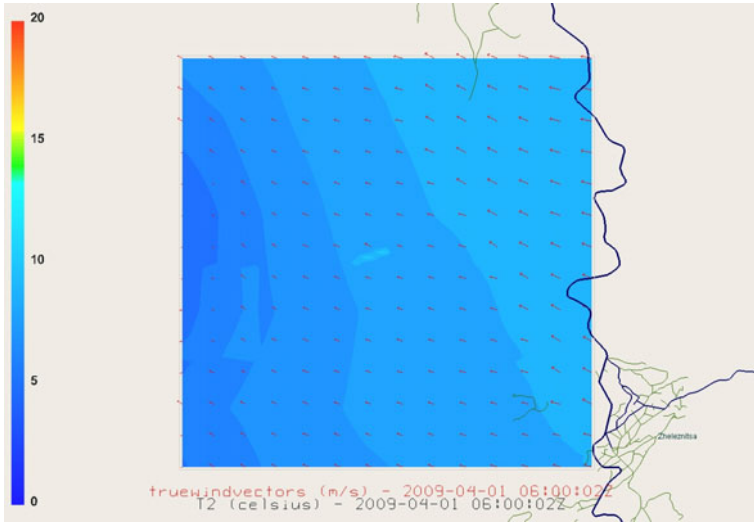


Fig. 1. Temperature (degrees C) at 2 m above the ground at the time of ignition, 2 s after the simulation start. The wind vectors are at 10 m height above the ground. The mean 10m wind speed is around 3 m/s. The ignition line is visible.

(2 seconds after the start of the simulation). The time step used in this simulation is 0.5 s. The boundary conditions are specified and are delivered from the WRF preprocessor WPS. The WRF physics parameterizations used are [14, Chapter 5]: Microphysics - Lin et al. scheme (`mp_physics` = 2), Longwave radiation - RRTM scheme (`ra_lw_physics` = 1), Shortwave Radiation - Dudhia scheme (`ra_sw_physics` = 1), Surface Layer - MM5 similarity (`sf_sfclay_physics` = 1), Land Surface - 5-layer thermal diffusion (`sf_surface_physics` = 1), Planetary Boundary layer - Yonsei University scheme (`bl_pbl_physics` = 1). Instead of real fuel data, the fuel used in the fire simulation is based on the altitude (`fire_fuel_read`=1). The large-scale meteorological background data has 1 degree horizontal resolution and is obtained from NCEP Global Analysis Data. The input for land cover and land use data is from the standard data sources of WRF obtained from USGS with 1km resolution with global coverage (<http://edc2.usgs.gov/glcc/glcc.php>). The terrain input is also from the standard WRF data sources, USGS with 1 km resolution.

Clearly, the resolution of the land cover and the terrain data is too coarse for realistic high-resolution studies, but this is the first step of the testing the WRF-Fire capabilities to work with real data in Bulgaria. For future experiments, more detailed data will be used. We are in the process of acquiring databases of high-resolution topography and databases of fuel data, which specify the type of trees (oak, or a type of conifer). We expect to specify the custom fuel categories available in WRF-Fire to input the data for Bulgarian mountains.

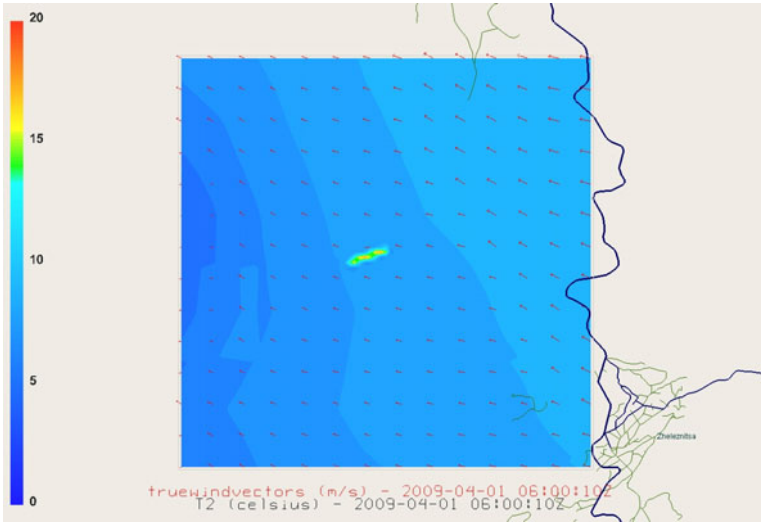


Fig. 2. The fire is strongly burning and it has spread in the wind direction, 10 s after the simulation start

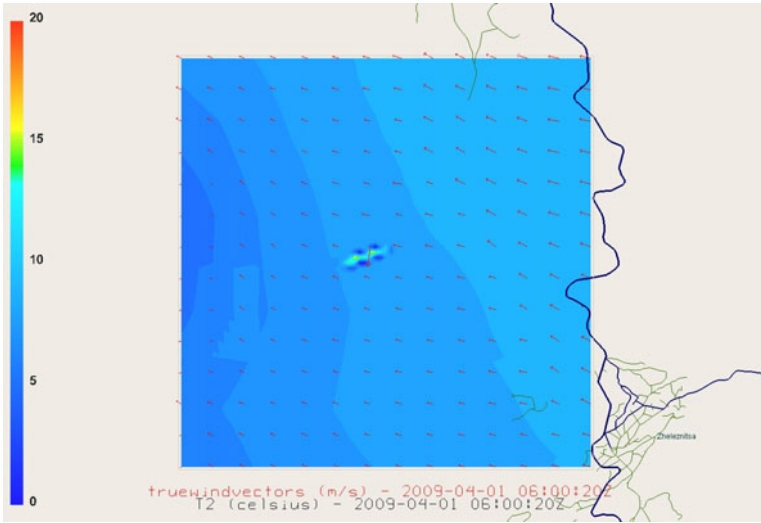


Fig. 3. The interior of the burning area is cooling down, 20 s after the simulation start

In Figs. 1 – 3, we have given the temperature change in the first stage of the fire propagation, the stage of the fire when the flame is very strongly burning, and the last picture is with fire intensity slowing down. The three pictures show also the wind direction and the fire propagation line. The simulation scenario is a real representation of possible forest fire at Vitosha mountain, 10 km south

from Sofia. The results from this experiment prove the capabilities of the model to work with real data for areas in Bulgaria and to give results suitable for forecasting the propagation of the fire line. The coupling of meteorological model with fire model gives us the abilities to take into account in our experiments the influence of the wind on the fire, and, conversely, the winds created by the fire itself. Obtaining as good meteorological and land data as possible is very important for getting adequate results in future real cases.

4 Conclusion and Future Plans

In this paper we have described how a wildland fire in Vitosha mountain can be simulated by adapting WRF-Fire v.3.2. Since we were limited by the real data we had and used idealizations and approximations for the fuel, the fire was not real, but it does approximate a possible real scenario. We have used raw data with approximations, because the real data is not available yet for our research.

Our future goal is to incorporate successfully real data from the wildland fire, which occurred near by the village Leshnikovo in the region of Haskovo, municipality of Harmanli in August 2009. For the new data, we plan to use 100 meters resolution land-use and land-cover data instead of the 1 km now, and high-resolution topography along with real fuel data. WRF-Fire is still an experimental tool for wildland modelling. Our team is aware that most of its applications are adapted for the U.S. territory and running it with Bulgarian data has added new features to the model specifics. We intend to make all results and successful runs available online for other WRF-Fire users.

Acknowledgements

This work was supported by the European Social Fund and Bulgarian Ministry of Education, Youth and Science under Operative Program “Human Resources Development,” Grant BG051PO001-3.3.04/40, the project DMU 02-14 “Collecting and Processing of Data Concerning Wild land Fires, Occurred on the Bulgarian Territory in the Recent Years by Using Weather Research and Forecasting Model-Fire (WRF-Fire)”, and by the U.S. National Science Foundation under grants AGS-0835579 and CNS-0719641.

References

1. Anderson, H.E.: Aids to determining fuel models for estimating fire behavior. USDA Forest Service, Intermountain Forest and Range Experiment Station, Research Report INT-122 (1982), http://www.fs.fed.us/rm/pubs_int/int_gtr122.html
2. Beezley, J.D.: How to run WRF-Fire with real data, http://www.openwfm.org/wiki/How_to_run_WRF-Fire_with_real_data (visited July 2010)

3. Clark, T.L., Coen, J., Latham, D.: Description of a coupled atmosphere-fire model. *International Journal of Wildland Fire* 13, 49–64 (2004)
4. Coen, J.L.: Simulation of the Big Elk Fire using coupled atmosphere-fire modeling. *International Journal of Wildland Fire* 14, 49–59 (2005)
5. Dobrinkova, N., Nedelchev, L.: PERUN – a system for early warning and simulation of forest fires and other natural or human-caused disasters. In: Adakin, E.E., Zakonnova, L.I., Vertchagina, I.Y., Dolganov, D.N. (eds.) *Education and science. Kemerovskii State University, VII International Scientific Conference “Nauka i obrazovanie”*, Belovo, Russia, March 14–15 (2008)
6. Editorial: Forest fires reach catastrophic scales (in Bulgarian). *Ecopolis* 48 (2001), http://www.bluelink.net/bg/bulletins/ecopolis12/1_os_1.htm
7. Finney, M.A., McHugh, C.W., Grenfell, I.C.: Stand- and landscape-level effects of prescribed burning on two Arizona wildfires. *Canadian Journal of Forest Research – Revue Canadienne de Recherche Forestiere* 35, 1714–1722 (2005)
8. Mandel, J., Chen, M., Franca, L.P., Johns, C., Puhalskii, A., Coen, J.L., Douglas, C.C., Kremens, R., Vodacek, A., Zhao, W.: A note on dynamic data driven wildfire modeling. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) *ICCS 2004. LNCS*, vol. 3038, pp. 725–731. Springer, Heidelberg (2004)
9. Mandel, J., Beezley, J.D., Coen, J.L., Kim, M.: Data assimilation for wildland fires: Ensemble Kalman filters in coupled atmosphere-surface models. *IEEE Control Systems Magazine* 29, 47–65 (2009)
10. National Fire Safety and Civil Protection Service of Bulgaria: Statistics – Forest fires (in Bulgarian), http://www.nspbn.mvr.bg/Sprav_informacia/Statistika/gorski.htm (visited July 2010)
11. Osher, S., Fedkiw, R.: *Level set methods and dynamic implicit surfaces*. Springer, New York (2003)
12. Patton, E.G., Coen, J.L.: WRF-Fire: A coupled atmosphere-fire module for WRF. In: *Preprints of Joint MM5/Weather Research and Forecasting Model Users’ Workshop*, Boulder, CO, June 22–25, pp. 221–223. NCAR (2004), http://www.mmm.ucar.edu/mm5/workshop/ws04/Session9/Patton_Edward.pdf
13. Sullivan, A.L.: A review of wildland fire spread modelling, 1990–present. *International Journal of WildLand Fire* 18, 347–403 (2009)
14. Wang, W., Bruyère, C., Duda, M., Dudhia, J., Gill, D., Lin, H.C., Michalakes, J., Rizvi, S., Zhang, X., Beezley, J.D., Coen, J.L., Mandel, J.: *ARW version 3 modeling system user’s guide. Mesoscale & Microscale Meteorology Division, National Center for Atmospheric Research* (July 2010), http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/ARWUsersGuideV3.pdf
15. WRF Working Group: *Weather Research Forecasting (WRF) Model*, <http://wrf-model.org> (visited July 2010)

Bulgarian Operative System for Chemical Weather Forecast

Iglika Etropolska¹, Maria Prodanova¹, Dimiter Syrakov¹, Kostadin Ganev²,
Nikolai Miloshev², and Kiril Slavov¹

¹ National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences,
Sofia, Bulgaria

`iglika.etropolska@meteo.bg`

² Geophysical Institute, Bulgarian Academy of Sciences, Sofia, Bulgaria
`kganev@geophys.bas.bg`

Abstract. In the paper, an operational prototype of the Integrated Bulgarian Chemical Weather Forecasting and Information System is presented. The system is foreseen to provide in real time forecast of the spatial/temporal Air Quality behavior for the country and (with higher resolution) for selected sub-regions and cities. The country-scale part of the system has been designed and tested and is now running operationally. It is based on the US EPA Models-3 System (MM5, SMOKE and CMAQ). The meteorological input to the system is NIMHs operational numerical weather forecast. The emission input exploits a high resolution disaggregation of the EMEP 50x50 km inventory for the year 2003. When elaborated, the actual national emission inventory is foreseen to be used. The boundary conditions are prepared by a similar system running operationally at Aristotle University of Thessaloniki, Greece. The System automatically runs twice a day (00 and 12 UTC) and produces 48-hour forecast. The results of each Systems run are post-processed in a way to archive the most important pollutants forecasts as to compare them with the respective measurements for the sake of verification of the System. Part of these pollutants is visualized as sequences of maps giving the evolution of the air quality over the country. The plots are uploaded to NIMHs web-server. The web-site is designed in a way to show both forecast maps for specified moments of time and animations for the entire 48-hour period for a number of key species.

1 Introduction

The Air Quality (AQ) is a key element for the well-being and quality of life of European citizens. According to the World Health Organization, air pollution severely affects the health of European citizens [19,20]: between 2.5 and 11% of the total number of annual deaths are due to air pollution. There is increasing evidence for adverse effects of air pollution on both the respiratory and the cardiovascular system as a result of both acute and chronic exposure. In particular, a significant reduction of life expectancy by a year or more is assumed to be linked to long-term exposure to high air concentrations of particulate matter

(PM). There is considerable concern about impaired and detrimental air quality conditions over many areas in Europe, especially in urbanized areas, in spite of about 30 years of legislation and emission reduction. Current legislation, e.g. the Ozone daughter directive 2002/3/EC [9], requires informing the public on AQ, assessing air pollutant concentrations throughout the whole territory of Member States and indicating exceedances of limit and target values, forecasting potential exceedances and assessing possible emergency measures to abate exceedances. For the purpose modeling tools must be used in parallel with air pollution measurements. The goals of reliable air quality forecasts are the efficient control and protection of population exposure as well as possible emission abatement measures. In last years, the concept of "chemical weather" arises and in many countries respective forecast systems are being developed along with the usual meteorological weather forecasts.

Air pollution easily crosses national borders. It would be cost-effective and beneficial for citizens, society and decision-makers that national chemical weather forecast and information systems would be networked across Europe. For the purpose a new COST Action ES0602 "Towards a European Network on Chemical Weather Forecasting and Information Systems" was launched aiming at providing a forum for harmonizing, standardizing and benchmarking approaches and practices in data exchange and multi-model capabilities for air quality forecast and (near) real-time information systems in Europe. It is supposed to examine existing and work out new solutions for integrating the development efforts at national and international levels. It will serve as a platform for the information exchange between the meteorological services, environmental agencies, and international initiatives. Detailed description of this COST Action can be found in [11,12]. There, one can find several CW systems description. Much more information about various European CW systems (performance and descriptions) is placed on the Action web-portal (<http://www.chemicalweather.eu/Domains>).

Bulgaria joined the COST Action ES0602 from its very beginning. This participation invoked a project supported by the National Science Fund with the Bulgarian Ministry of Education, Youth and Science. Its main purpose was to create the Bulgarian Chemical Weather forecast and information system (BGCW) intended to provide timely informative and reliable forecasts tailored to the needs of various users. Here, the current level to which the BGCW has been developed is described and some results and end-user products are provided.

2 Short Description of BGCW Structure and Information Flow

The country-size part of BGCW is designed in a way to fit the real-time constraints and to deliver forecasts twice a day (00 and 12 UTC) for the next 48 hours. US EPA Models-3 air quality modeling system is used, here, consisting of:

- CMAQ (<http://www.cmaq-model.org/>), Community Multi-scale Air Quality model being the chemical-transport model (CTM) of the System [2,3,6]

- MM5 (<http://box.mmm.ucar.edu/mm5/>), The 5th generation PSU/NCAR Meso-meteorological Model used as meteorological pre-processor to CMAQ [7,10], and
- SMOKE (<http://www.smoke-model.org/>), Sparse Matrix Operator Kernel Emissions modelling system being the emission pre-processor to CMAQ [4,5]

Meteorological forecasts are obtained at the main synoptic terms by ALADIN-model being Bulgarian national weather forecast tool. The computational domain covers Bulgaria with d resolution of 10 km and ALADIN output is transformed to GRIB format with 6-hour time resolution.

In Fig. 1, the data flow diagram for one 48-hour cycle is displayed. In the boxes, together with the names of the systems elements, the format of the respective output is given. The white boxes present Models-3 components, the brown (dark grey) ones the created interface modules (FORTRAN codes). The green (light grey) boxes present the data input to the system. First of all, this is the meteorological forecast created by ALADIN which drives MM5. The MM5 vertical structure consists of 23 -levels with varying thickness, extending up to 100 hPa height. Proper physical options are set to MM5. The FDDA option [16] is switched on keeping MM5s forecast close to the ALADINs one. MM5 starts its calculation 12 hours earlier for spin-up reasons (i.e. MM5 performs 60-hour run).

MCIP, the Meteorology-Chemistry Interface Processor, is part of CMAQ and together with the needed meteorological parameters prepares some other data (fluxes, dry deposition velocities etc.) to be used by CMAQ and SMOKE. Area Source (AS) gridded inventory feeds the AEmis (AS emission processor). The Large Point Source (LPS) inventory is input to SMOKE together with the ambient meteorological data as to produce LPS emissions (LPS processor). The met-data together with the gridded land-use are used by SMOKE to produce biogenic emissions (BgS processor). SMOKE is used once more to merge these 3 emission files in a model-ready emission input.

Apart these two main inputs meteorology and emissions CMAQ needs initial and boundary conditions. The initial conditions are taken from previous CMAQ run. The case with the boundary conditions (BC) is much more complicated. BCs are of great importance for small regions like Bulgaria. In the current version of BGCW, the boundary conditions are provided by the chemical weather forecast system running operationally in Aristotle University of Thessaloniki (AUTH), Greece [13,14]. AUTH system exploits the nested domain approach. The air quality forecast is carried out for Europe (50 km spatial resolution), the Balkans (10km) and Athens (2km) using the photochemical air quality model CAMx [8]. AUTH system is run once a day producing 3-day pollution forecast. The 3-day real-time boundary conditions for CMAQ are prepared in two steps. First, CAMx data is interpolated for CMAQ boundary points, an operational procedure that takes place in AUTH (module CW.BC1). Its results are uploaded to a dedicated sever in Sofia. Here, this data is processed on-line (module CW.BC2) as to produce 3-day CMAQ-ready boundary condition file (Fig. 1).

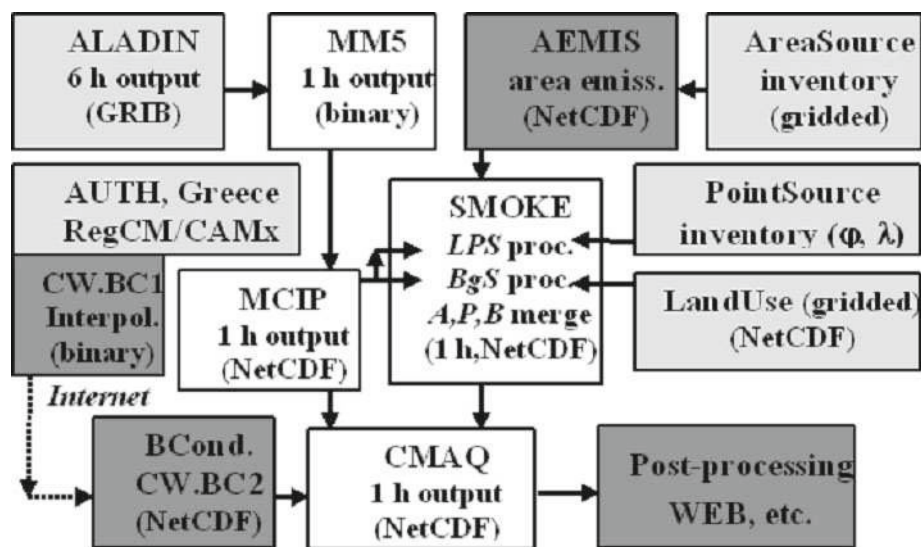


Fig. 1. Data flow of a 48-hour BGCW forecast

CMAQ demands its emission input in specific format reflecting the time evolution of all polluters accounted for by the used chemical mechanism. Emission inventories are made on annual basis for big territories and many pollutants are estimated as groups. As to prepare emission input to CTM, gridding, time allocation (monthly, weekly, daily, profiles provided by [1]) and speciation (splitting the group pollutants) must take place. Emission models are needed for the purpose. Such a component in Models-3 system is SMOKE. As already mentioned, it is partly used, here, only for calculating LPS and BgS emissions and to merge AS-, LPS- and BgS-files. Input to these interfaces is gridded emission inventory of TNO [18] for 2003. This inventory is a disaggregation of the 50-km EMEP inventory and contains data for 10 source types (SNAPs). The speciation profiles are elaborated on the base of US EPA ones (<http://www.epa.gov/ttn/chief/emch/speciation/>) and are used by both AEmiss and SMOKE. The biogenic emissions are prepared by SMOKE by the BEIS-3.13 mechanism [15] on the base of gridded land-use (USGS 24 categories).

More detailed description of BGCW, as well as its validation is given in [17].

3 Operational Performance of BGCW

Fourteen σ -levels with varying thickness determine the vertical structure of CMAQ. The Planetary Boundary Layer (PBL) is presented by 8 of these levels. The daily CMAQ, v.4.6., output is a NetCDF file with 3D hourly data for 78 pollutants, from which: 52 gaseous, 21 aerosols (Aitken and accumulation modes), 5 aerosol distributions (3 by number, 2 by aerosol area). The last box in Fig. 1 tags the post-processing activity that is quite important as to BGCW

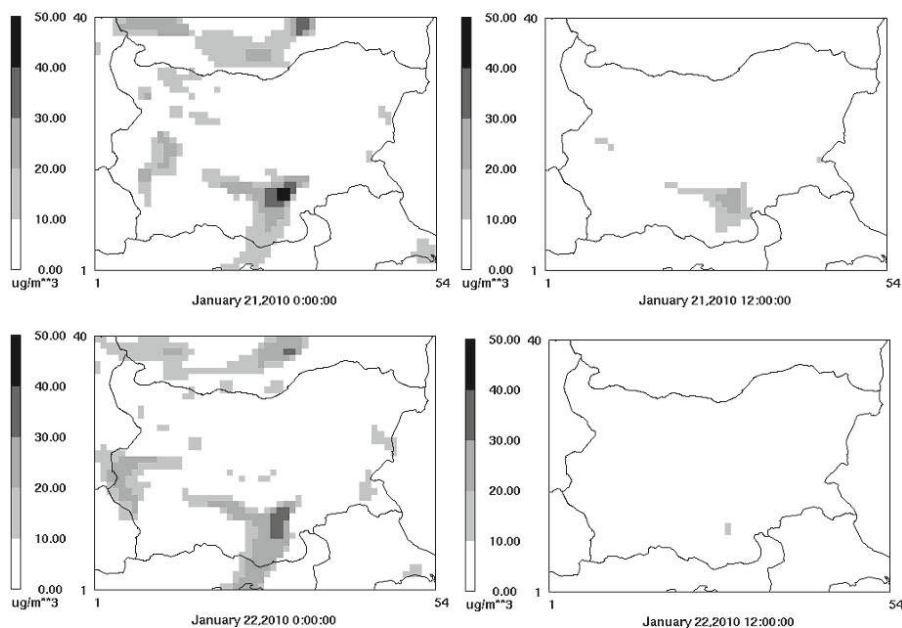


Fig. 2. Forecasted time evolution of NO_2 starting from January 21, 2010, 00:00 UTC

results become visual. First of all the post-processing program XtrCON extracts part of the pollutants for archiving and further handling. Only surface values of the most important 17 pollutants are saved – 8 gases and 9 aerosol types. Part of these pollutants is more or less monitored and they are referred in the legislations with the respective thresholds. It must be mentioned that the sum of all aerosol compounds forms PM_{10} (usually measured) and $\text{PM}_{2.5} = \text{PM}_{10} - \text{CPRM}$, CPRM being the coarse mode aerosol. All this data is stored in a single NetCDF file with the current Julian date as file-name. In Fig. 2, the forecasted evolution of NO_2 is displayed with time resolution of 12 hour.

One can notice the specific spatial behavior of this pollutant: the area of the regions with higher concentrations decreases during the day-time hours and increases during the night. The maximal values for the region (about $50 \mu\text{g}/\text{cub.m}$) appear in the cells near to "Mariza-Iztok" Thermal Power Plants (TPPs), a set of three lignite coal burning TPPs, being the most powerful SO_2 , NO_2 and PM polluters in the Balkans. The diurnal variation of NO_2 in one of these cells is shown in Fig. 3 and it confirms this fact.

The reason for this behavior is possibly the diurnal variation of the turbulence in PBL. Although it is a winter period, during the day the sun warms the lower atmospheric layers enough to decrease the PBL stability. The generated turbulence takes up the pollution and the concentrations at the ground decrease compared to the night-time hours. The same can be noticed in Fig. 4, where the forecasted evolution of PM_{10} concentrations for the first 12 hours of the period is presented.

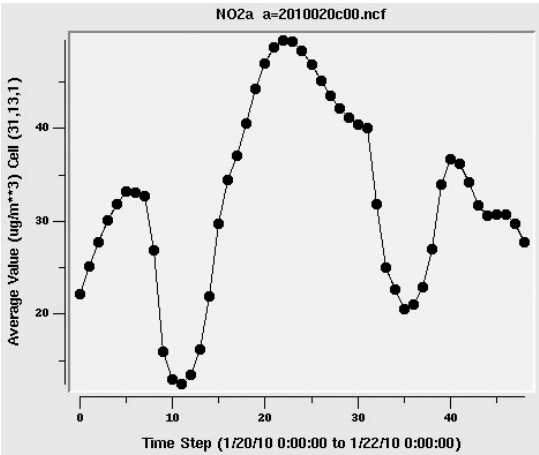


Fig. 3. Surface NO_2 time profile in the vicinity of TPP "Mariza-Iztok"

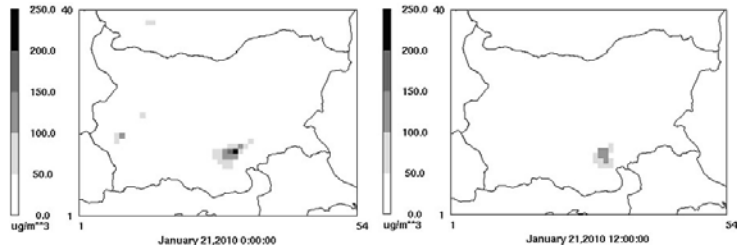


Fig. 4. Forecasted time evolution of PM10

The most prominent feature of the period is the well expressed plumes from the big polluters, the main of which is TPP "Mariza-Iztok". The period is characterized by a cold invasion from North-North-East as it can be seen on the plots. The PM10 diurnal variation is similar to that of NO_2 but is not expressed so well.

As to make the results of BGCW operation public, a specialized web-site was created on the NIMH server (<http://info.meteo.bg/cw/frameset.html>). For the moment it presents 4 main pollutants Ozone, NO_2 , SO_2 and PM10 (Fig. 5). Every pollutant is invoked by clicking in the list at the left side of the page. All 48 hourly forecast fields can be retrieved by putting the cursor on the respective line in the scale at the right side of the graph. Putting the cursor over "Play"-line invokes animation of the forecast. Complementary to ozone hourly fields the absolute and 8-hour averaged maxima for the first and the second day can be visualized. Under each pollutants view respective local thresholds are displayed. The title of the page is linked to a pdf-file with full description of Bulgarian CW forecast system.

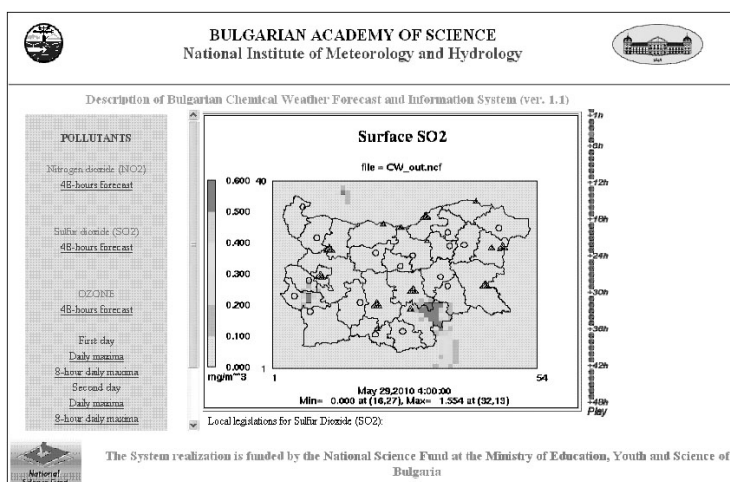


Fig. 5. View of the Bulgarian Chemical weather forecast web-site

4 Conclusion

The country-scale part of Bulgarian Chemical Weather Forecast and Information System is designed on the base of US EPA Models-3 System: MM5 (meteorological pre-processor), SMOKE (emission pre-processor) and CMAQ (Chemical Transport Model). The meteorological input to the system is the ALADIN output, ALADIN being the national numerical weather forecast tool. At this stage, the emission input exploits the high resolution inventory for year 2003 produced by TNO, The Netherlands, the actual national emission inventory is foreseen to be used. The boundary conditions are prepared by a similar system running operationally in Aristotle University of Thessaloniki, Greece. The System automatically runs twice a day (00 and 12 UTC) and produces 48-hour forecast.

The evaluation of BGCW simulations showed that the System has a satisfactory performance with respect to O₃, despite using boundary conditions from another modeling system. The best simulation quality refers to summer time daily maximums. The reasonable performance of the System for the hindcast simulations (see [17]) justifies its use for future forecast and information provided to various users. At the moment, the system is running automatically twice a day. The results of each Systems run are post-processed in a way to archive the most important pollutants. Part of these pollutants is visualized as sequences of maps giving the evolution of the air quality over Bulgaria and can be seen on the respective web-site (<http://info.meteo.bg/cw/frameset.html>).

Acknowledgments

This study is made under the financial support of Bulgarian National Science Fund (Grant No. D002-161/16.12.2008). The presented results were not possible

without the experience obtained during the participation in the FP5 project BULAIR (Contract No. EVK2-CT-2002-80024), FP6 Network of Excellence ACCENT (Contract No. GOCE-CT-2002-500337) and the FP6 Integrated Project QUANTIFY (Contract No. 003893 GOG). The contacts within the framework of the NATO SFP Project No. 981382 were extremely stimulating as well.

Deep gratitude is due to all organizations providing free of charge data and software used in this study, namely US EPA, US NCEP and European institutions like EMEP, EEA and many others. Special thanks to the Netherlands Organization for Applied Scientific research (TNO) for providing us with the high-resolution European anthropogenic emission inventory and emission time allocation profiles.

References

1. Builtjes, P.J.H., van Loon, M., Schaap, M., Teeuwisse, S., Visschedijk, A.J.H., Bloos, J.P.: Project on the modelling and verification of ozone reduction strategies: contribution of TNO-MEP, TNO-report, MEP-R2003/166, Apeldoorn, The Netherlands (2003)
2. Byun, D., Ching, J.: Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. EPA Report 600/R-99/030, Washington DC (1999)
3. Byun, D., Schere, K.L.: Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. *Applied Mechanics Reviews* 59(2), 51–77 (2006)
4. CEP: Sparse Matrix Operator Kernel Emission (SMOKE) Modeling System, University of North Carolina, Carolina Environmental Programs - CEP, Research Triangle Park, North Carolina (2003)
5. Coats Jr., C.J., Houyoux, M.R.: Fast Emissions Modeling With the Sparse Matrix Operator Kernel Emissions Modeling System. In: *The Emissions Inventory: Key to Planning, Permits, Compliance, and Reporting*, Air and Waste Management Association, New Orleans (September 1996)
6. Dennis, R.L., Byun, D.W., Novak, J.H., Galluppi, K.J., Coats, C.J., Vouk, A.: The Next Generation of Integrated Air Quality Modeling: EPA's Models-3. *Atmosph. Environment* 30, 1925–1938 (1996)
7. Dudhia, J.: A non-hydrostatic version of the Penn State/NCAR Mesoscale Model: validation tests and simulation of an Atlantic cyclone and cold front. *Monthly Weather Review* 121, 1493–1513 (1993)
8. ENVIRON: Users Guide to the Comprehensive Air Quality Model with Extensions (CAMx), Version 4.40, ENVIRON International Corporation, Novato, CA (2006)
9. European Parliament: DIRECTIVE 2002/3/EC of 12 February 2002 relating to ozone in ambient air. *Official Journal of the European Communities* (9.3.2002), L67, 14–30 (2002)
10. Grell, G.A., Dudhia, J., Stauffer, D.R.: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR Technical Note, NCAR/TN-398+STR (1994)
11. Karatzas, K., Kukkonen, J. (eds.): COST Action ES0602 Quality of life information services towards a sustainable society for the atmospheric environment. COST Office (2009) ISBN: 978-960-6706-20-2

12. Kukkonen, J., Klein, T., Karatzas, K., Torseth, K., Fahre Vik, A., San Jose, R., Balk, T., Sofiev, M.: COST ES0602: towards a European network on chemical weather forecasting and information systems. *Advances in Science and Research* 3, 27–33 (2009)
13. Poupkou, A., Kioutsioukis, I., Lisaridis, I., Markakis, K., Giannaros, T., Katragkou, E., Melas, D., Zerefos, C., Viras, L.: Evaluation in the Greater Athens Area of an air quality forecast system. In: *Proc. of the IX EMTE National-International Conference of Meteorology-Climatology and Atmospheric Physics*, Thessaloniki, Greece, May 28-31, pp. 759–766 (2008a)
14. Poupkou, A., Kioutsioukis, I., Lisaridis, I., Markakis, K., Melas, D., Zerefos, C., Giannaros, T.: Air quality forecasting for Europe, the Balkans and Athens. In: *3rd Environmental Conference of Macedonia*, Thessaloniki, Greece, March 14-17 (2008b)
15. Schwede, D., Pouliot, G., Pierce, T.: Changes to the Biogenic Emissions Inventory System Version 3 (BEIS3). In: *Proc. of 4th Annual CMAS Models-3 Users's Conference*, Chapel Hill, NC, September 26-28 (2005)
16. Stauffer, D.R., Seaman, N.L.: Use of four-dimensional data assimilation in a limited area mesoscale model. Part I: experiments with synoptic data. *Monthly Weather Review* 118, 1250–1277 (1990)
17. Syrakov, D., Ganev, K., Prodanova, M., Slavov, K., Etropolska, I., Miloshev, N., Jordanov, G.: Background pollution forecast over Bulgaria. In: *18th International Symposium ECOLOGY & SAFETY*, Sunny Beach, Bulgaria, June 8 - 12, vol. 3, Part 1, pp. 32–41 (2009) (published on a CD), (Published at: ISSN: 1313-2563), <http://www.science-journals.eu>
18. Visschedijk, A.J.H., Zandveld, P.Y.J., Denier van der Gon, H.A.C.: A High Resolution Gridded European Emission Database for the EU Integrate Project GEMS, TNO-report 2007-A-R0233/B, Apeldoorn, The Netherlands (2007)
19. WHO: Fact Sheet Number 187, World Health Organization (2000)
20. WHO: Health aspects of air pollution: Results from the WHO project Systematic review of health aspects of air pollution in Europe. WHO Regional Office For Europe, Copenhagen, Denmark (2004), <http://www.euro.who.int/document/E83080.pdf>

Atmospheric Composition Studies for the Balkan Region

Georgi Gadzhev¹, Georgi Jordanov¹, Kostadin Ganev¹, Maria Prodanova²,
Dimitir Syrakov², and Nikolai Miloshev¹

¹ Geophysical Institute, Bulgarian Academy of Sciences, Sofia, Bulgaria
`kganev@geophys.bas.bg`

² National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences,
Sofia, Bulgaria
`dsyrak@eagle.cu.bas.bg`

Abstract. The present work aims at studying the local to regional atmospheric pollution transport and transformation processes over the Balkan Peninsula and at tracking and characterizing the main pathways and processes that lead to atmospheric composition formation in the region.

The US EPA Models-3 system is chosen as a modelling tool, its nesting capabilities applied for downscaling the simulations to a 9 km resolution over Balkans. The TNO emission inventory is used as emission input. Special pre-processing procedures are created for introducing temporal profiles and speciation of the emissions.

The Models-3 "Integrated Process Rate Analysis" option is applied to discriminate the role of different dynamic and chemical processes for the pollution from road and ship transport. Some results from several emission scenarios which make it possible to evaluate the contribution of different SNAP are demonstrated as well.

1 Introduction

Regional studies of the air pollution over the Balkans, including country-to-country pollution exchange, had been carried out for quite a long time - see for example [12,7,8,15,4,10,11,14]. These studies were focused on both studying some specific air pollution episodes and long-term simulations and produced valuable knowledge and experience about the regional to local processes that form the air pollution pattern over Southeast Europe. It seems, however, that the impact of the different SNAP categories to the air pollution of the Balkan Peninsula as a whole had never been comprehensively studied. Carrying out such a study with up-to-date modelling tools detailed and reliable input data for long enough simulation periods and good resolution is the aim of the present work.

The air pollution pattern is formed as a result of interaction of different processes, so knowing the contribution of each for different meteorological conditions and given emission spatial configuration and temporal behaviour is by all means important. That is why the present study attempts to make some evaluations of the contribution of different processes to the local to regional pollution over the Balkans.

In order to obtain high quality scientifically robust assessments of the air quality and its origin it is clear that extensive sensitivity studies have to be carried out. The grid computing technology ([1,6,8]), is applied for the purpose.

2 Metodology, Models and Input Data

The US EPA Model-3 system was chosen as a modelling tool. The system consists of three components: MM5 - the 5th generation PSU/NCAR Meso-meteorological Model MM5 - [5,9], used as meteorological pre-processor; CMAQ - the Community Multiscale Air Quality System CMAQ - [3,2]; SMOKE - the Sparse Matrix Operator Kernel Emissions Modelling System. The large scale (background) meteorological data used by the application is the NCEP Global Analysis Data with $1^\circ \times 1^\circ$ resolution. At the moment the created database contains all the necessary information since year 2000.

The TNO high resolution inventory ([13]) is exploited. The inventory is produced by proper disaggregation of the EMEP 50-km inventory data base. The TNO inventory resolution is $0.125^\circ \times 0.625^\circ$ longitude-latitude, that is on average about $14 \times 7 \text{ km}$. GIS technology is applied as to produce area and large point source input from this data base. It must be mentioned that the TNO emissions are distributed over 10 SNAPs (Selected Nomenclature for Air Pollution) classifying pollution sources according the processes leading to harmful material release to the atmosphere.

CMAQ demands its emission input in specific format reflecting the time evolution of all polluters accounted for the used chemical mechanism. A specific approach for obtaining speciation profiles is used here. The USA EPA data base is intensively exploited. A Bulgarian emission expert has found coincidence between main Bulgarian sources for every SNAP with similar source types from US EPA nomenclature. The weighted averages of the respective speciation profiles are accepted as SNAP-specific splitting factors, weights being the percentage of contribution of every source type in total Bulgarian emission in particular SNAP. In such a way VOC and PM_{2.5} speciation profiles are derived.

As far as the background meteorological data is the NCEP Global Analysis Data with $1^\circ \times 1^\circ$ resolution, it is necessary to use MM5 and CMAQ nesting capabilities as to downscale to 3 km step for the innermost domain. The MM5 pre-processing program TERRAIN was used to define four domains with 81 (D1), 27 (D2), 9 (D3) and 3 (D4) km horizontal resolution. These four nested domains were chosen in such a way that the domain with a horizontal resolution of 9 km covers the whole Balkan Peninsula.

In order to evaluate the contribution of the emissions from different SNAP categories several emission scenarios were run: all emissions included (basic scenario) and emissions from chosen SNAP category reduced by a factor of 0.8, which makes it possible to calculate the relative contribution of the respective SNAP category to the overall pollution.

The Models-3 "Integrated Process Rate Analysis" option is applied to discriminate the role of different dynamic and chemical processes for the air pollution

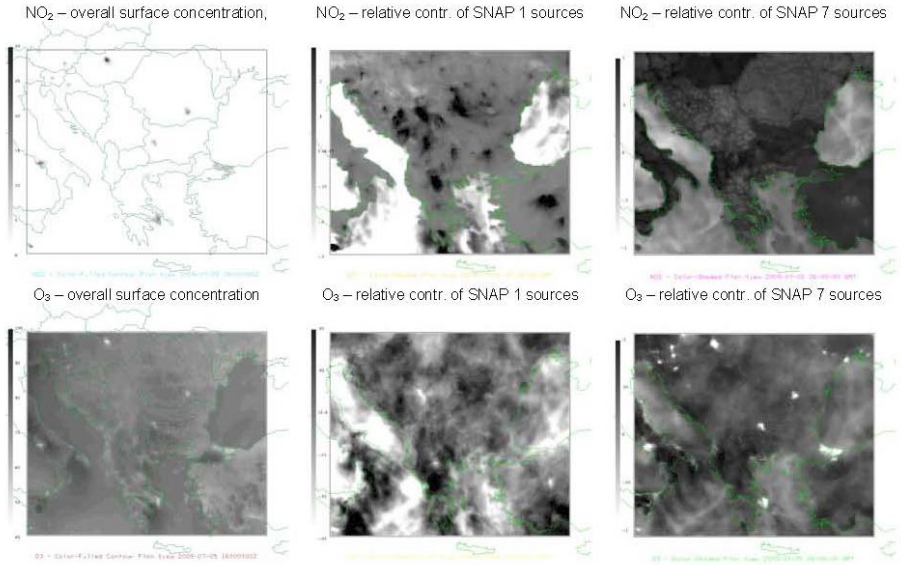


Fig. 1. Plots of surface concentrations [$\mu\text{g}/\text{m}^3$] and relative contributions of SNAP 1 and 7 sources for NO_2 and O_3 for 16 UTC, "typical" day in July

pattern formation. The procedure allows the concentration change for each compound for an hour Δ_c to be presented as a sum of the contributions Δc_i of the processes, which determine the concentration. The processes that are considered are: advection, diffusion, mass adjustment, emissions, dry deposition, chemistry, aerosol processes and cloud processes/aqueous chemistry.

The modelling infrastructure (models and input data, Grid simulation practices) has been well validated (see for example [8] which allows applying it for air pollution studies for the Balkan region with some trust in the obtained results. MM5 and CMAQ simulations were carried out for the years 2003–2009 and the respective pollution characteristics for each day for all the period were cPlots of surface concentrations [$\mu\text{g}/\text{m}^3$] and relative contributions of SNAP 1 and 7 sources for NO_2 and O_3 for 16 UTC, "typical" day in July, calculated. Averaging the daily fields over the whole ensemble of results for the respective month produces a diurnal behaviour of given pollution characteristic, which can be interpreted as "typical" for the month (respectively season).

3 Some Examples of the SNAP Categories Contribution Simulations

The characteristic, which will be demonstrated and discussed as an example further in this paragraph is the surface concentration c for months January and July. Analyzing the results one should keep in mind that the contribution of a

SNAP category in a given point (or sub-domain) reflects the respective SNAP sources in the whole integration domain.

Plots of surface concentrations of O_3 and NO_2 , typical for July, for 16 UTC are shown in Figure 1, together with the relative contribution of the emissions from SNAP categories 1 and 7. One can not help, but notice how the big industrial sites, big cities, roads are manifested as sources in the NO_2 plots and as sinks in the O_3 plots. The plot of the SNAP 7 relative contribution to NO_2 pollution can almost be used as a map of the roads in the region.

It should be noted that the emission contribution could be negative. That is not such a surprise, having in mind that the atmospheric compounds are subjects to very complex and none-linear chemical transformations.

Plots of this kind are rather spectacular and can give a good qualitative impression of the spatial complexity of the different SNAP categories contribution. In order to demonstrate the pollution and SNAP code contribution behavior in a more simple and easy to comprehend way, the respective fields can be averaged over some domain (in this case the territory of Bulgaria), which makes it possible to jointly follow and compare the diurnal behavior of the overall pollution and the pollution from the respective SNAP categories (obtained by multiplying the relative contribution by the concentration from all the sources).

Such plots for some of the compounds are given in Figure 2 for January and. There are several things in these plots which should be mentioned. First of all the contribution of both SNAP categories to the NO_2 , PM2.5 and PM-coarse is quite significant (the SNAP 7 contribution for NO_2 is definitely bigger than the SNAP 1 one). The diurnal courses of the PM2.5 and PM-coarse overall concentrations, as well as the concentrations from Snap categories 1 and 7 for January and July are very similar.

The overall surface concentrations of NO_2 , PM2.5 and PM-coarse have similar specifics for the overall concentration and the part of the concentration due to SNAP 7 emissions peaks in the morning and late afternoon, early evening hours and a minimum around noon. This obviously is due partially to the specific diurnal course of the road transport (SNAP 7) emissions, but probably also to the meteorological conditions a tendency for predominant unstable conditions during the day, which causes more intensive vertical mixing, thus transporting some of the surface pollution aloft.

The O_3 diurnal course manifests the expected maximum in daytime. Surprisingly the contribution of both SNAP categories to the overall O_3 concentration is small. This probably means that the O_3 in Bulgaria is mostly "imported", while the Bulgarian sources cause O_3 formation somewhere else. This is not so revolutionary conclusion, having in mind that the O_3 is a secondary pollutant and can be formed away from the O_3 precursor sources.

The more general features, characteristic for January are valid for July as well. In July the atmospheric stability effect can be followed also in the diurnal course of the concentrations due to SNAP 1 emissions (elevated sources) slight peaks around noon can be observed for NO_2 , PM2.5 and PM-coarse.

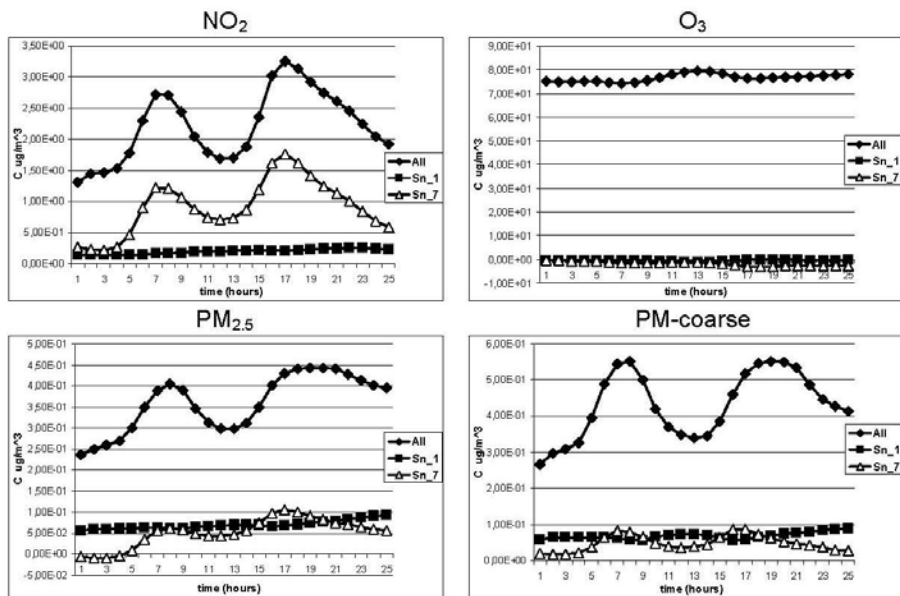


Fig. 2. Plots of the diurnal course of surface concentrations [$\mu\text{g}/\text{m}^3$] from all the sources and from the sources from SNAP 1 and 7 sources for NO_2 , O_3 , $\text{PM}_{2.5}$ and PM-coarse for a "typical" day in January

4 Some Examples of Process Analysis Simulations

The characteristic, which will be discussed as an example further in this paragraph are the surface process contributions Δc_i and the resulting hourly surface concentration changes Δc for January. Due to the limited volume of the paper no surface 2D plots of the process contributions will be demonstrated. Very briefly, it can be said that their pattern is indeed very complex, but some typical effects can be followed, for example the roads, big cities and agglomerations appear as big sinks in the O_3 chemical transformation plots and as big sources in the $\text{PM}_{2.5}$ aerosol processes plots.

Plots for NO_2 , SO_2 averaged over Bulgaria for January are given in Figure 3. A detailed description even of these simple images will take a lot of space and probably is not necessary. Some more general features could be mentioned (valid for summer months as well), however: i) The hourly surface concentration Δc is determined mainly by a small number of most important processes (which could be different for different compounds), while the role of the others is minor; ii) The temporal behavior of the processes is also complex; iii) For some processes the contribution sign is obvious (like emissions or dry deposition), but some can change their sign during the day; iv) For all of the compounds some of the advection/diffusion processes have a major role.

As can be seen from Figure 2 the road transport is a major source for NO_2 and so the resulting surface concentration Δc (empty circles) follows the course of

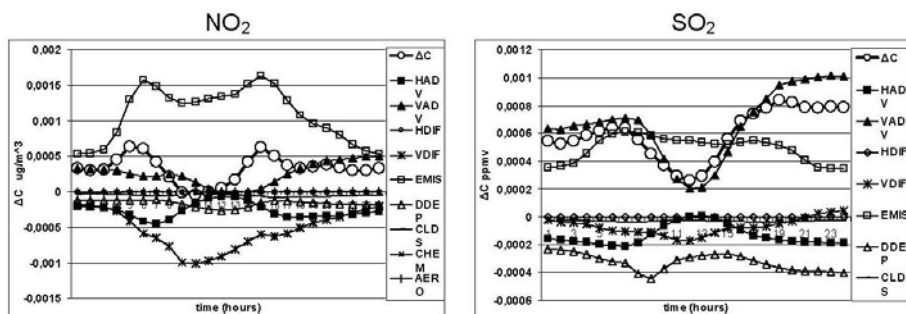


Fig. 3. Plots of the diurnal course of the contributions [$\mu\text{g}/\text{m}^3$ and ppmV] of the different processes to the formation of NO_2 and SO_2 and the resulting hourly concentration change Δc for a "typical" day in January

the emissions (empty squares). The SO_2 on the other hand is to large extend due to elevated sources (power plants, etc.). That is why the surface concentration change follows the vertical transport (black triangles). It could be also noted that the horizontal advection (black squares) acts oppositely to the vertical advection. That is quite easy to comprehend, having in mind the continuity law. The major SO_2 sink appears to be the dry deposition, while for the NO_2 it is the vertical diffusion which transports the near surface generated NO_2 aloft.

5 Conclusion

The numerical experiments performed produced a huge volume of information, which have to be carefully analyzed and generalized so that some final conclusions could be made. The obtained ensemble of numerical simulation results is extensive enough to allow statistical treatment – calculating not only the mean concentrations and different SNAP categories contribution mean fields, but also standard deviations, skewness, etc. with their dominant temporal modes (seasonal and/or diurnal variations).

The results produced by the CMAQ "Integrated Process Rate Analysis" demonstrate the very complex behaviour and interaction of the different processes – process contributions change very quickly with time and these changes for the different points on the plane hardly correlate at all. The analysis of the behaviour of different processes does not give simple answer of the question how the air pollution in a given point or region is formed.

Acknowledgments

The present work is supported by the project SEE-GRID-SCI - contract No. FP7 RI-211338, as well as by the Bulgarian National Science Fund (grants No. 002-161/2008 and 002-115/2008).

Deep gratitude is due to US EPA, US NCEP and EMEP for providing free-of-charge data and software. Special thanks to the Netherlands Organization for Applied Scientific research (TNO) for providing high-resolution European anthropogenic emission inventory.

The work of the young scientists G. Jordanov and G. Gadzhev is also supported by the ESF project No. BG51PO001-3.3.04-33/28.08.2009. G. Gadzhev is World Federation of Scientists grant holder.

References

1. Atanasov, E., Gurov, T., Karaivanova, A.: Computational Grid: structure and Applications. *Journal Avtomatika i Informatica*, 40–43 (September 2006) (in Bulgarian), ISSN 0861-7562, 3/2006, year XL
2. Byun, D., Ching, J.: Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. EPA Report 600/R-99/030, Washington DC (1999)
3. Byun, D., Young, J., Gipson, G., Godowitch, J., Binkowski, F.S., Roselle, S., Benjey, B., Pleim, J., Ching, J., Novak, J., Coats, C., Odman, T., Hanna, A., Alapaty, K., Mathur, R., McHenry, J., Shankar, U., Fine, S., Xiu, A., Jang, C.: Description of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. In: 10th Joint Conference on the Applications of Air Pollution Meteorology with the A& WMA, Phoenix, Arizona, January 11–16, pp. 264–268 (1998)
4. Chervenkov, H.: Estimation of the Exchange of Sulphur Pollution in the Southeast. *Europe Journal of Environmental Protection and Ecology* 7(1), 10–18 (2006)
5. Dudhia, J.: A non-hydrostatic version of the Penn State/NCAR Mesoscale Model: validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.* 121, 1493–1513 (1993)
6. Foster, J., Kesselmann, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1998)
7. Ganev, K., Dimitrova, R., Syrakov, D., Zerefos, C.: Accounting for the mesoscale effects on the air pollution in some cases of large sulfur pollution in Bulgaria or Northern Greece. *Environmental Fluid Mechanics* 3, 41–53 (2003)
8. Ganev, K., Syrakov, D., Prodanova, M., Miloshev, N., Jordanov, G., Gadzhev, G., Todorova, A.: Atmospheric composition modeling for the Balkan region. In: SEE-GRID-SCI USER FORUM 2009, Istanbul, December 9–10, pp. 77–87 (2009) ISBN: 978-975-403-510-0
9. Grell, G.A., Dudhia, J., Stauffer, D.R.: A description of the Fifth Generation Penn State/NCAR Mesoscale Model (MM5). NCAR Technical Note, NCAR TN-398-STR, 138 pp (1994)
10. Poupkou, A., Symeonidis, P., Lisaridis, I., Melas, D., Ziomas, I., Yay, O.D., Balis, D.: Effects of anthropogenic emission sources on maximum ozone concentrations over Greece. *Atmospheric Research* 89(4), 374–381 (2008)
11. Symeonidis, P., Poupkou, A., Gkantou, A., Melas, D., Devrim Yay, O., Pouspourika, E., Balis, D.: Development of a computational system for estimating biogenic NMVOCs emissions based on GIS technology. *Atmospheric Environment* 42(8), 1777–1789 (2008)
12. Syrakov, D., Prodanova, M., Ganev, K., Zerefos, C., Vasaras, A.: Exchange of sulfur pollution between Bulgaria and Greece. *Environmental Science and Pollution Research* 9(5), 321–326 (2002)

13. Visschedijk, A., Zandveld, P., van der Gon, H.: A high resolution gridded European emission database for the EU integrated project GEMS, TNO report 2007-A-R0233/B (2007)
14. Zanis, P., Poupkou, A., Amiridis, V., Melas, D., Mihalopoulos, N., Zerefos, C., Katragkou, E., Markakis, K.: Effects on surface atmospheric photo-oxidants over Greece during the total solar eclipse event of 29 March 2006. *Atmospheric Chemistry and Physics Discussions* 7(4), 11399–11428 (2007)
15. Zerefos, C., Ganey, K., Kourtidis, K., Tzortziou, M., Vasaras, A., Syrakov, E.: On the origin of SO₂ above Northern Greece. *Geophysical Research Letters* 27(3), 365–368 (2000)

Specialized Sparse Matrices Solver in the Chemical Part of an Environmental Model

Krassimir Georgiev¹ and Zahari Zlatev²

¹ Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
georgiev@parallel.bas.bg

² National Environmental Research Institute,
Aarhus University Frederiksborgvej 399,
P.O. Box 358, DK-4000 Roskilde, Denmark
zz@dmu.dk

Abstract. A two-dimensional advection-diffusion-chemistry module of a large-scale environmental model (Danish Eulerian Model for studying the transport of air pollutants on large scale - UNI-DEM) is taken. The module is described mathematically by system of partial differential equations. Sequential splitting is used in the numerical treatment. The non-linear chemistry is most the time-consuming part during the computer runs and it is handled by six implicit algorithms for solving ordinary differential equations. This leads to the solution of very long sequences of systems of linear algebraic equations. It is crucial to solve these systems efficiently. This is achieved by applying four different algorithms which are developed, tested and discussed.

Keywords: Environmental models, Advection-diffusion-chemistry module, Partial differential equations, Ordinary differential equations, Systems of linear algebraic equations, Sparse techniques.

1 The 2D Version of the Danish Eulerian Model and Rotational Test

The Danish Eulerian model (UNI-DEM) ([16]) is a model for studying the long-range transport of air pollutants. The model computational domain covers Europe and parts of Asia, Africa and the Atlantic Ocean. The long-range transport of air pollution is usually studied by a system of partial differential equations (PDEs), which can be written as follows (it should be mentioned that similar systems are used in other environmental models):

$$\begin{aligned} \frac{\partial c_s}{\partial t} = & -\frac{\partial(uc_s)}{\partial x} - \frac{\partial(vc_s)}{\partial y} \\ & + \frac{\partial}{\partial x} \left(K_x \frac{\partial c_s}{\partial x} \right) + \frac{\partial}{\partial y} \left(K_y \frac{\partial c_s}{\partial y} \right) \\ & + E_s + Q_s - (k_{1s} + k_{2s})c_s, \quad s = 1, 2, \dots, q, \end{aligned} \quad (1)$$

where:

- (i) $c_s(t, x, y)$ are the concentrations of the chemical species;
- (ii) $u(t, x, y)$, $v(t, x, y)$ are the wind components along the coordinate axes;
- (iii) $K_x(t, x, y)$, $K_y(t, x, y)$ are the diffusion coefficients;
- (iv) $E_s(t, x, y)$ present the emission sources;
- (v) $k_{1s}(t, x, y)$, $k_{2s}(t, x, y)$ are correspondingly the dry and wet deposition coefficients, and finally,
- (vi) $Q_s(t, x, y, c_1, c_2, \dots, c_q)$ are non-linear functions which describe the chemical reactions between species under consideration.

When some numerical methods should be tested and tuned it is more convenient to use the following two-dimensional module:

$$\begin{aligned} \frac{\partial c_s}{\partial t} = & -\mu(y - y_0) \frac{\partial c_s}{\partial x} - \mu(x_0 - x) \frac{\partial c_s}{\partial y} + K \left(\frac{\partial c_s}{\partial x^2} + \frac{\partial c_s}{\partial y^2} \right) \\ & + E_s(t, x, y) + Q_s(t, x, y, c_1, c_2, \dots, c_q), \quad s = 1, 2, \dots, q, \end{aligned} \quad (2)$$

where $x \in [a_1, b_1]$, $y \in [a_2, b_2]$, $t \in [a, b]$ $x_0 = (b_1 - a_1)/2$, $y_0 = (b_2 - a_2)/2$ and $\mu = 2\pi/(b - a)$. Very often the space domain is square, i.e. $a_1 = a_2, b_1 = b_2$ and the length of the time interval is 86 400 s. (24 hours with starting point being 6:00 AM). The problem (2) is considered with a given initial value vector $c(a, x, y)$ and some boundary conditions (Dirichlet boundary conditions which be used hereafter).

The most essential difference in (2) according to (1) is the special definition of the wind velocity field. The trajectories of the wind in (2) are concentric circles with center (x_0, y_0) and particles are rotated along these trajectories with a constant angular velocity. Such wind velocity field was first defined in [4,11].

The non-linear terms Q_s , $s = 1, 2, \dots, q$ describe the chemical reactions (these are precisely the same as those in one of the chemical schemes used in UNI-DEM). Some of these reactions are photo-chemical. The photo-chemical reactions are deactivated during the night and activated during the day. Therefore, the length of the time-interval should be at least 24 hours in order to study the performance during the changes from day-time to night-time and from night-time to day-time when some of the concentrations are very rapidly changing.

Most often the problem (2) is run in the following two typical cases: (a) when all emissions E_s , $s = 1, 2, \dots, q$ are set equal to zero (*puff*) and (b) when there are some non-zero emissions (*plume*). In the latter case we have: (i) all non-zero emissions are specified in a circle with centre $(x_0, y_0) = (0.25(b_1 - a_1), 0.5(b_2 - a_2))$ and radius $r = 0.125(b_1 - a_1)$, (ii) the emissions outside this circle are equal to zero, (iii) the emissions form a cone, the highest emissions being in the centre (x_0, y_0) of the circle.

There, the test in which only the first two terms in the right-hand-side of (2) are kept (the pure advection test) was introduced. Chemical reactions were included to the Crowley-Molenkamp test by Hov et al (see [8]). The module defined in this paper by (2) is a further extension of the original Crowley-Molenkamp

test. It is worthwhile to study this module (generalized rotation test) because it is much closer to real environmental models than the previous two modules. By setting some of the coefficients to zero or keeping all of them different from zero, six situations can be studied by (2):

- **No non-zero emissions are specified** (*puff*; $E_s = 0$ for all values of s):
 - (A) Pure advection-diffusion process (only the terms containing the spatial derivatives are kept).
 - (B) Pure chemical process (only the non-linear functions are kept).
 - (C) Combining the advection-diffusion process with the chemistry process (all terms except the emissions are kept).
- **Some emissions are no zero** (*plume*; $E_s \neq 0$ for some values of s):
 - (D) Pure advection-diffusion process (only the terms containing the spatial derivatives and the emissions are kept).
 - (E) Pure chemical process (only the non-linear functions Q_s and the emissions are kept).
 - (F) Combining the advection-diffusion process with the chemistry process (all terms are kept).

2 Sequential Splitting Procedure

A splitting procedure proposed in [10] is used in the Danish Eulerian model. In the two dimensional version of the model it leads to four submodels representing the horizontal advection, the horizontal diffusion, the chemistry and the emissions, and the depositions (dry and wet). Simple sequential splitting will be used in this paper. Applying this kind of splitting to (2) leads to the following two sub-problems:

$$\frac{\partial g_s}{\partial t} = -\mu(y - y_0) \frac{\partial g_s}{\partial x} - \mu(x_0 - x) \frac{\partial g_s}{\partial y} + K \left(\frac{\partial^2 g_s}{\partial x^2} + \frac{\partial^2 g_s}{\partial y^2} \right), \quad (3)$$

$$\frac{\partial h_s}{\partial t} = E_s(t, x, y) + Q_s(t, x, y, g_1, g_2, \dots, g_q) \quad (4)$$

Assume that the time-integration is carried out by using a constant stepsize Δt and that some approximation $\bar{c}_i(t_n, x, y)$ to the exact solution $c_i(t_n, x, y)$ of (3) at $t_n = a + n\Delta t$ has been calculated. Then $\bar{g}_i(t_n, x, y)$ is set equal to $\bar{c}_i(t_n, x, y)$ and an approximation $\bar{g}_i(t_{n+1}, x, y)$ to the exact solution $g_i(t_{n+1}, x, y)$ of (4) at $t_{n+1} = a + (n+1)\Delta t$ is computed by using an appropriate numerical method. The second sub-problem is handled in a similar way: $\bar{h}_i(t_n, x, y)$ is set equal to $\bar{g}_i(t_{n+1}, x, y)$ and an approximation $\bar{h}_i(t_{n+1}, x, y)$ to the exact solution $h_i(t_{n+1}, x, y)$ of (4) at $t_{n+1} = a + (n+1)\Delta t$ is computed by using an appropriate numerical method. Finally, $\bar{c}_i(t_{n+1}, x, y)$ is set equal to $\bar{h}_i(t_{n+1}, x, y)$, which completes the computations at an arbitrary time-step $n = 1$. The computations start with the initial vector $c(t_0, x, y) = c(a, x, y)$ which is given in advance.

It is worthwhile to emphasize here that the following two remarks are important when a sequential splitting procedure is used:

- The sequential splitting procedure allows different numerical methods to be used in the treatment of the two sub-problems. This is a very useful feature, because the two sub-problems have different properties. The first sub-problem, the advection-diffusion module is a non-stiff problem, while the second sub-problem, the chemistry module is a stiff sub-problem.
- The system (3) consists of q independent PDEs. If the computational space domain is discretized into $N_x \times N_y$ grid-points, then (4) will be decoupled into $N_x \times N_y$ independent systems of ordinary differential equations (ODEs) each of which contains q equations. This observation indicates that in general efficient parallel computations can be achieved in a natural way.

3 Numerical Treatment in the Chemistry Module

The chemical submodel is described by (4). Let the space domain is discretized by using $(N_x + 1)(N_y + 1)$ grid points. Then the non-linear system of PDEs (4) is reduced to $(N_x + 1)(N_y + 1)$ independent systems of ODEs (one per each grid-point in the space domain). The number of equations in every system of ODEs is q . An arbitrary system of ODEs can be written as

$$\frac{\partial \hat{h}}{\partial t} = f(t, \hat{h}), \quad \hat{h} \in \mathcal{R}^q, \quad f \in \mathcal{R}^q. \quad (5)$$

Assume that system (5) is obtained at the grid-point (x_i, y_j) where $i = 0, 1, \dots, N_x$ and $j = 0, 1, \dots, N_y$. It is clear that the components $\hat{h}_k$, $k = 1, 2, \dots, q$ of vector $\hat{h}$ are approximations of the concentrations at grid-point (x_i, y_j) and at time t , while the k th component of vector f can be written in the following way:

$$f_k(t, \hat{h}) = E_k(t, x_i, y_j) + Q_k(t, x_i, y_j, \hat{h}_1, \hat{h}_2, \dots, \hat{h}_k) \quad k = 1, 2, \dots, q. \quad (6)$$

The system of ODEs (5) is stiff. Moreover, the treatment of this system is very time-consuming. Therefore, it is crucial to achieve fast computational process in the solution of (5). Three tools can be used to resolve efficiently this task: (a) selecting fast numerical methods for solving systems of stiff systems of ODEs, (b) developing efficient methods for solving systems of linear algebraic equations and (c) parallelizing the computations. The treatment of (b) and (c) will be discussed in the following section. Six methods for solving systems of stiff ODEs have been tried in the attempt to resolve (a). These methods as well as their accuracy and stability properties are listed in Table 1. It should be mentioned that L-stability is stronger than A-stability (the class of the L-stable methods is a sub-class of the class of A-stable methods).

Let us denote with $J(t, \hat{h}) = \frac{\partial f(t, \hat{h})}{\partial \hat{h}} \in \mathcal{R}^{q \times q}$ the Jacobian matrix of the vector function $f \in \mathcal{R}^q$ (the right-hand-side of (5)). Let $\hat{h}_{old}$ be an approximation to $\hat{h}$ calculated by any of the six numerical methods listed in Table 1 at $t = t_n$. Since

Table 1. Computing times measure in CPU hours for different space and time discretizations

Numerical method	Order	Stability properties	Reference
Backward Euler Method	first	L-stable	[2,5,9]
<i>Implicit Mid-point Rule</i>	<i>second</i>	<i>A-stable</i>	[2,6,9]
Two stage Modified Diagonally	second	L-stable	[14,15]
Implicit Runge-Kutta Method			
<i>Three-stage Fully Implicit</i>	<i>fifth</i>	<i>L-stable</i>	[2,6]
<i>Runge-Kutta Method</i>			
Two-stage Rosenbrock Method	second	L-stable	[6,7]
<i>Trapezoidal Rule</i>	<i>second</i>	<i>A-stable</i>	[2,6,9]

all six methods in Table 1 are implicit and the system of ODEs (5) is non-linear, it is necessary to apply the Newton Iterative Method (see [3]) in order to find an approximation $\hat{h}_{new}$ to $\hat{h}$ at $t = t_{n+1}$. Let $\hat{h}_{new}^{[m]}$ be the vector obtained at the m th iteration. At each iteration m of the Newton Iterative Method, the following system of linear algebraic equations has to be solved:

$$B_m z_m = d_m, \quad \hat{h}_{new}^{[m+1]} = \hat{h}_{new}^{[m]} + z_m, \tag{7}$$

where

$$B_m = I - \gamma \Delta t J \left(t_{n+1}, \hat{h}_{new}^{[m+1]} \right), \quad z_m = \Delta \hat{h}_{new}^{[m+1]}, \tag{8}$$

$$d_m = s \left(\hat{h}_{old}, \hat{h}_{new}^{[m]} \right), \quad m = 0, 1, 2, \dots \tag{9}$$

Here, I is the identity matrix in $\mathcal{R}^{q \times q}$. Note that (8) and (9) can be used with all numerical methods from Table 1 (except the Three-stage Fully Implicit Runge-Kutta Method). Parameter γ is in general varies with the method. For example, $\gamma = 1$ when the Backward Euler Method is used, while $\gamma = 0.5$ when the Trapezoidal Rule is selected.

The situation becomes more complicated when the Three-stage Fully Implicit Runge-Kutta Method is applied. However, (8) and (9) can still be used provided that $I \in \mathcal{R}^{3q \times 3q}$ and γJ is replaced by a matrix $\hat{J} \in \mathcal{R}^{3q \times 3q}$ (matrix $\hat{J}$ can be partitioned into a 3×3 -block-matrix, each block containing matrix J multiplied by a scalar). This shows that the difference is only quantitative (the system solved is much bigger when the Three-stage Fully Implicit Runge-Kutta Method is selected).

The above analysis shows that an important part of the computational work is the solution of systems of linear algebraic equations (9). The number of these systems per time-step is $(N_x + 1)(N_y + 1)$ for all numerical methods in Table 1. The number of equations per system is q for all methods except the Three-stage Fully Implicit Runge-Kutta Method; for the latter method the number is $3q$. Note that $(N_x + 1)(N_y + 1)$ is normally very large (up to many millions), while

q is normally small; $q \in [20, 200]$. Therefore, it is very important to find an answer to the following questions: how can a very large number of small systems of linear algebraic equations be solved efficiently? An answer to this important question will be given in the next section.

4 Numerical Treatment of the Sparse Matrices Arising in the Chemistry Modle

The properties and characteristics of the matrices arising in the chemistry module which are listed bellow are essential:

- they are **general** (these matrices **ARE NOT** symmetric, diagonally dominant, banded or positive definite);
- they are **badly scaled**;
- they are **ill-conditioned**;
- their **elements vary in large intervals** on diurnally basis which cause great diurnal variation of the involved concentrations (an example, which illustrates the diurnal variation of the hydroxyl radical is given in Fig. 1).

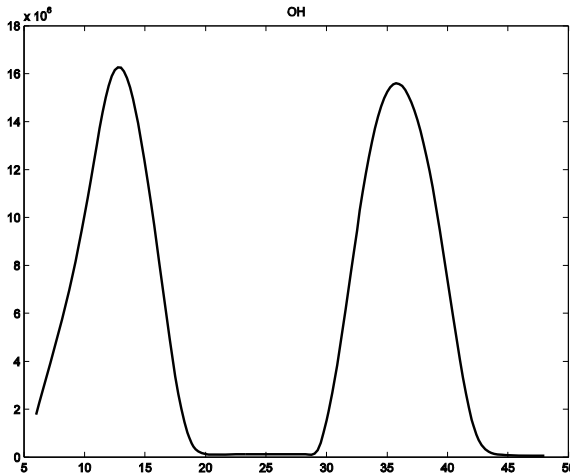


Fig. 1. Diurnal variation of the hydroxyl radical

The following several techniques for treatment of the matrices which arise in the chemistry module are used in the performed numerical and computer experiments:

- **Dense matrix technique:** *LAPACK* – (DGETRF and DGETRS) (see e.g. in [1]);
- **Regular sparse matrix technique:** *PARASPAR* – DIR (Direct solution using Gaussian Elimination) (see e.g. in [13,15]);

Table 2. Computing times measured in CPU hours for different space and time discretizations

Method	Time-steps	$N_x \times N_y$	Dense	Sparse direct	Precond. sparse	Special sparse
Backward	960	1089	0.112	0.067	0.052	0.040
Euler	1920	4225	0.621	0.269	0.157	0.063
Method	3840	16641	4.249	1.732	0.875	0.241
(first order)	7680	66049	31.383	12.575	6.238	1.506
	15360	263169	250.712	98.921	52.775	12.116
Implicit	960	1089	0.243	0.118	0.087	0.039
Midpoint	1920	4225	1.507	0.608	0.360	0.059
Rule	3840	16641	9.464	3.693	2.030	0.200
(second order)	7680	66049	61.622	22.930	11.772	1.407
	15360	263169	384.331	138.737	72.002	11.647
Two-stage	960	1089	0.205	0.088	0.064	0.041
2nd order	1920	4225	1.400	0.402	0.234	0.071
diag.implicit	3840	16641	9.944	2.891	1.552	0.269
Runge-Kutta	7680	66049	75.052	22.665	11.880	1.972
Method	15360	263169	542.686	179.742	93.748	15.479
3-stage	960	1089	0.301	0.322	0.410	0.128
5th order	1920	4225	1.958	2.108	1.152	0.668
fully impl.	3840	16641	13.862	15.184	13.294	4.543
Rrkge-Kutta	7680	66049	107.426	118.113	68.278	33.052
Method	15360	263169	657.868	921.914	403.838	268.724
Two-stage	960	1089	0.060	0.067	0.065	0.047
2-nd order	1920	4225	0.321	0.288	0.256	0.095
Rosenbrock	3840	16641	1.620	2.082	1.724	0.437
Method	7680	66049	12.939	16.976	13.070	2.801
	15360	263169	105.220	144.371	99.395	21.083
Trapezoidal	960	1089	0.118	0.063	0.051	0.039
Rule	1920	4225	0.632	0.233	0.145	0.056
(second order)	3840	16641	4.481	1.514	0.817	0.207
	7680	66049	33.966	11.445	6.198	1.155
	15360	263169	252.372	91.759	49.137	10.369

- **Preconditioned sparse matrix technique:** *PARASPAR* – ORTH (Preconditioned Modified Orthomin) (see e.g. in [12,15]);
- **Special sparse matrix technique** – only for the chemical scheme in UNI-DEM.

The main problems which appear when a regular sparse matrix code is applied for small matrices, are related mainly to: (i) *the indirect addressing*, (ii) *the use of many integer arrays* the content of which must very often be updated, (iii) *the performance of many short loops* and (iv) *the need to search for pivotal elements at every stage of the Gaussian elimination* (in an attempt to preserve both the

sparsity and and the stability). Therefore, the special sparse matrix technique has been developed. This technique is based on the following steps:

- *A preliminary reordering procedure* based on the application of a Markovitz pivotal strategy for general sparse matrices is performed. All small non-zero elements are removed during this preliminary procedure.
- The positions, in which new *non-zero elements* will be created, are determined and locations for these elements are reserved in the arrays where the LU factorization of the diagonal block under consideration is stored.
- *A loop-free code* for the numerical calculation of the *LU factorization* of the diagonal block under consideration is prepared.
- *A loop-free code for the back-substitution* (based on the LU factorization computed in the previous step) is prepared.

5 Numerical Results

The computer experiments were carried out in parallel using OpenMP tools to parallelize the code. The parallel code is fully described in [17]. The computing time when six different numerical methods for solving systems of ODEs and four different techniques for treatment of the sparse matrices arising in the chemical model discussed above are presented in Table 2.

The meaning of the column 4 - 7 in Tabl. 2 are as follows:

- **Dense technique:** the matrices of the systems the linear algebraic equations are treated as dense matrices and LAPACK subroutines are called to factorize them and to solve the systems.
- **Sparse direct:** traditional sparse matrix technique is used; the systems are solved directly and the option for solving systems of linear algebraic equations directly, which is available in package PARASPAR, is used.
- **Preconditioned sparse:** an approximate LU factorization is calculated by dropping small elements and used as a preconditioner. The modified ORTHOMIN method from package PARASPAR is used. When the method is not convergent, the drop-tolerance is reduced in an attempt to calculate more accurate preconditioner.
- **Special sparse:** a special sparse matrix technique, suitable only for the particular chemical scheme used in UNI-DEM, is applied.

6 Conclusion

An important component of the large-scale air pollution models, a chemistry module, was studied in this paper. The numerical treatment of this module requires the solution of very long sequences of systems of linear algebraic equations. It has been shown in this paper that the selection of efficient algorithms for solving systems of linear algebraic equations is very important. The computing time can be reduced by a large factor (sometimes by a factor larger than 30) when the proper method is chosen.

Acknowledgments

This research is supported in part by grants DO02-115/2008 (CVP_09_002) and DO 02-147/2008 from the Bulgarian NSF and the Bulgarian Supercomputer Center (NCSA) giving access to IBM Blue Gene/P computer.

References

1. Alexandrov, V., Sameh, A., Siddique, Y., Zlatev, Z.: Numerical integration of chemical ODE problems arising in air pollution models. *Environmental Modelling and Assessment* 2, 365–377 (1997)
2. Butcher, J.C.: *The numerical analysis of ordinary differential equations. Runge-Kutta methods and general linear methods.* Wiley, Chichester (1987)
3. Golub, G.H., Ortega, J.M.: *Scientific computing and differential equations.* Academic Press, London (1992)
4. Crowley, W.P.: Numerical advection experiments. *Monthly Weather Review* 96, 1–11 (1968)
5. Hairer, E., Nørsett, S.P., Wanner, G.: *Solving ordinary differential equations, I: Nonstiff problems.* Springer, Heidelberg (1987)
6. Hairer, E., Wanner, G.: *Solving ordinary differential equations, II: Stiff and differential-algebraic problems.* Springer, Berlin (1991)
7. Hundsdorfer, W., Verwer, J.G.: *Numerical solution of time-dependent advection-diffusion-reaction equations.* Springer, Berlin (2003)
8. Hov, Ø., Zlatev, Z., Berkowicz, R., Eliassen, A., Prahm, L.P.: Comparison of numerical techniques for use in air pollution models with non-linear chemical reactions. *Atmospheric Environment* 23, 967–983 (1988)
9. Lambert, J.D.: *Numerical methods for ordinary differential equations.* Wiley, Chichester (1991)
10. Marchuk, G.I.: *Mathematical modeling for the problem of the environment. Studies in Mathematics and Applications, vol. 16.* North-Holland, Amsterdam (1985)
11. Molenkamp, C.R.: Accuracy of finite-difference methods applied to the advection equation. *Journal of Applied Meteorology* 7, 160–167 (1968)
12. Vinsome, P.: ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear equations. In: *Proc. Fourth Sympos. on Reservoir Simulation, Society of Petr. Eng. of AIME* (1976)
13. Zlatev, Z.: On some pivotal strategies in Gaussian elimination by sparse technique. *SIAM Journal on Numerical Analysis* 17, 18–30 (1980)
14. Zlatev, Z.: Modified diagonally implicit Runge-Kutta methods. *SIAM Journal on Scientific and Statistical Computing* 2, 321–334 (1981)
15. Zlatev, Z.: *Computational methods for general sparse matrices.* Kluwer Academic Publishers, Dordrecht (1991)
16. Zlatev, Z.: *Computer treatment of large air pollution models.* Kluwer, Dordrecht (1995)
17. Zlatev, Z., Dimov, I.: *Computational and Environmental Challenges in Environmental Modelling.* Elsevier, Amsterdam (2006)

A Numerical Investigation for the Optimal Contaminant Inlet Positions in Horizontal Subsurface Flow Wetlands

Konstantinos Liolios¹, Vassilios Tsihrintzis¹, and Stefan Radev²

¹ Laboratory of Ecological Engineering and Technology, Department of Environmental Engineering, Democritus University of Thrace, GR-67100 Xanthi, Greece

kliolios@env.duth.gr

² Bulgarian Academy of Sciences, Institute of Mechanics, Acad. G. Bonchev Str., Bl. 4, 1113 Sofia, Bulgaria

stradev@imbm.bas.bg

Abstract. The paper presents a numerical treatment of flow and contaminant removal in porous media with emphasis to horizontal subsurface flow constructed wetlands. The purpose here is to find their optimal design characteristics as concerns the contaminant inlet positions, in order to maximize the removal efficiency.

Keywords: Computational Fluid Mechanics, Groundwater Flow, Contaminant Transport and Removal, Constructed Wetlands.

1 Introduction

Constructed wetlands (CW) are recently a good alternative solution for small settlements in order to treat municipal wastewater, see e.g. [1]–[3] and especially [4]. The use of these systems is currently becoming very popular in many countries. So it seems necessary to find their optimal design characteristics in order to maximize the removal efficiency and keep their area and construction cost to a minimum.

A numerical simulation of such systems is presented here, with emphasis to horizontal subsurface constructed wetlands. First the mathematical modeling of the problem is presented. Next, for the numerical simulation, the Visual MODFLOW code family is used. MODFLOW is a computational model based on the finite difference method and used for the simulation of groundwater flow and mass transport. Further, the numerical procedure is applied for the simulation of pilot-scale units of horizontal subsurface flow wetlands. The above pilot-scale units were constructed and operated in the Laboratory of Ecological Engineering and Technology of the Department of Environmental Engineering, in Xanthi, of Democritus University of Thrace (DUTH) [1,3]. The objective here is to study numerically the removal of Biochemical Oxygen Demand (BOD) in a unit with medium gravel and reeds (MG-R). For the needs of the present study, a control

of the simulation of the processes in the wetlands was done by comparison to existing experimental results. Finally, the effects of the inlet recharge positions, in order to obtain the optimum contaminant removal, have been numerically investigated.

2 Method of Analysis

2.1 The Mathematical Modeling

The partial differential equation describing the fate and transport of contaminants of species k in 3-D, transient groundwater flow systems can be written as follows [5,6]:

$$\frac{\partial(\theta C^k)}{\partial t} = \frac{\partial}{\partial x_i} \left(\theta D_{ij} \frac{\partial C^k}{\partial x_j} \right) - \frac{\partial}{\partial x_i} (\theta v_i C^k) + q_v C_s^k + \sum R_n \quad (1)$$

where: θ = porosity of the subsurface medium; C^k = dissolved concentration of species k , in $[ML^{-3}]$; D_{ij} = hydrodynamic dispersion coefficient tensor, in $[L^2T^{-1}]$; v_i = seepage or linear pore water velocity, in $[LT^{-1}]$; it is related to the specific discharge or Darcy flux through the relationship, $v_i = q_i \theta$; q_v = volumetric flow rate per unit volume of aquifer representing fluid, sources (positive) and sinks (negative), in $[T^{-1}]$; C_s^k = concentration of the source or sink flux for species k , in $[ML^{-3}]$; $\sum R_n$ = chemical reaction term, in $[ML^{-3}T^{-1}]$. This last term, in the simplest linear case, with no absorption, depends on the first-order removal coefficient λ and is equal to $(-\lambda C)$.

The above equation (1) is the governing equation underlying in the transport model. This transport and decay equation is linked to the flow equation through the Darcy relationship:

$$v_i = -\frac{K_{ij}}{\theta} \frac{\partial h}{\partial x_j} \quad (2)$$

where: K_{ij} = a component of the hydraulic conductivity tensor, in $[LT^{-1}]$; $h = h(x, y, z, t)$ = hydraulic head, in $[L]$.

The hydraulic head is obtained from the solution of the three-dimensional groundwater flow equation:

$$\frac{\partial}{\partial x_i} \left(K_i \frac{\partial h}{\partial x_j} \right) + q = S_s \frac{\partial h}{\partial t} \quad (3)$$

where: S = the specific storage of the porous materials, in $[L^{-1}]$; and q = the volumetric flow rate per unit area of aquifer representing fluid sources (positive) and sinks (negative), in $[T^{-1}]$.

So, the above equations (1)-(3), combined with appropriate initial and boundary conditions, describe the 3-dimensional flow of groundwater and the transport and decay of contaminants in a heterogeneous and anisotropic medium. Thus, for the case of one only ($k=1$) pollutant species, the unknowns of the problem are the following five space-time functions: the hydraulic head $h = h(x, y, z, t)$, the velocity $v_i = v_i(x, y, z, t)$, and the concentration $c = c(x, y, z, t)$.

2.2 The Numerical Simulation

The numerical solution of the problem can be obtained by using a suitable numerical method, see e.g [6,7,8,11]. Here, because the constructed wetlands have usually a rectangular scheme, the Finite Difference Method is chosen. This method is the basis for the code family MODFLOW accompanied by an effective computer package [6,7,8,9,10,11]. MODFLOW is a computational model based on finite difference method and used for the simulation of groundwater flow and mass transport. So the program MT3DMS of MODFLOW is used in the present paper for the analysis of the pilot-scale units, constructed and operated in DUTH, Xanthi, Greece. For more details see [13].

3 Applications to Pilot - Scale Wetland Units

3.1 Pilot-Scale Units Description

Five similar pilot-scale horizontal subsurface flow constructed wetlands (CW) have been constructed and are in operation in the Laboratory of Ecological Engineering and Technology, Department of Environmental Engineering of DUTH in Xanthi, Greece. They are rectangular tanks made of steel, with dimensions $L = 3$ m long, 0.75 m wide and 1 m deep. A schematic view of the experimental layout is shown in Figure 1. The wetland units are equipped with inlet and outlet hydraulic structures, similar to those used in real systems. These five pilot-scale units were operated continuously from January 2004 until January 2006 in parallel experiments, in order to investigate the effect of temperature, hydraulic residence time (HRT), vegetation type and porous media material and grain size on the performance of horizontal subsurface flow constructed wetlands treating wastewater. For details, see [1,3].

Three of the above five units contained medium gravel (MG) obtained from a quarry. The other two units contained fine gravel the one and cobbles the other one, both obtained from a river bed. The two units with MG were planted: one with common reeds (*R, Phragmites australis*) and one with cattails (*C, Typha latifolia*). The third MG one was kept unplanted for comparison reasons. The other two units were planted with common reeds. Planting and porous media combinations were appropriate for comparing the effects of vegetation and media type on the function of the system. Synthetic wastewater was introduced in the units. For more details concerning the experimental procedures and results, see Akkratos [3].

Based on the above experimental results, first in [12,13], by applying sensitivity analysis, a calibration of MODFLOW has been realized. Next, in [13], by applying inverse problem procedures [14], the optimal range of values for the first order reaction rate λ for BOD removal has been estimated. This rate is relevant to the chemical reaction term ΣR_n in (1) and depends on temperature.

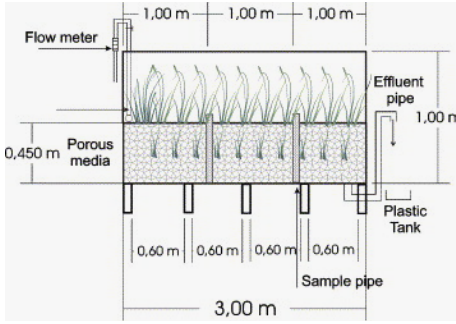


Fig. 1. Schematic section along one wetland tank

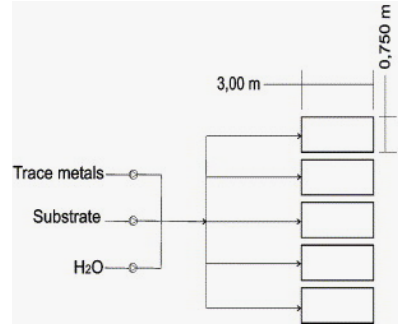


Fig. 2. General layout of the facility (plan)

3.2 The Investigated Unit MG-R with Reeds

Among the previously described five scale-units CW, the tank MG-R, containing medium gravel and reeds, has been here chosen indicatively for numerical investigation. The proposed numerical approach, based on MODFLOW and combined with MT3DMS module, has been applied [13].

Using available literature and laboratory estimates for various hydraulic and mass transfer parameters, see e.g. [15,16], and proper boundary and initial conditions, we evaluate the BOD concentration at selected points in the tank MG-R. So, for this tank, with length $L = 3$ m, the following input data are used: Confined aquifer storage: $S_s = 10^{-5} 1/m$; Unconfined aquifer storage: $S_y = 0.37$; Effective and total porosity: $\text{Eff.Por.} = \text{Tot.Por.} = 0.37$; Hydraulic conductivity: $K = 0.345 m/s$; Diffusion coefficient: $\text{Diff.Coeff.} = 0.0000036 m^2/hr$; Longitudinal dispersion: $\text{Long. Disp.} = 0.027$ m; Horizontal / Longitudinal dispersion = Vertical / Longitudinal dispersion = 0.01566 ; Initial hydraulic head: 0.45 m; Initial concentrations field: zero. Finally, the first-order reaction rate $\lambda = RC1$ for BOD removal estimated, as mentioned previously in [13], has a value-range $0.10 - 0.20$ ($days$) $^{-1}$.

In order to investigate the effects of the inlet recharge positions to the optimum contaminant removal, the following scenarios concerning the inlet places of the pollutant have been treated:

Case 1: Inlet of the whole (100%) pollutant at the entrance of the tank: $x_{in} = 0.0m$

Case 2: Inlet of the 60% of pollutant at the entrance of the tank and 40% at the place: $x_1 = 1m(L/3)$.

Case 3: Inlet of the 60% of pollutant at the entrance of the tank, 25% at the place $x_1 = 1m(L/3)$ and 15% at the place $x_2 = 2m(2L/3)$.

Case 4: Inlet of the 80% of pollutant at the entrance of the tank and 20% at the place: $x_1 = 1m(L/3)$.

Case 5: Inlet of the 90% of pollutant at the entrance of the tank and 10% at the place: $x_1 = 1m(L/3)$.

Representative results of the so obtained numerical ones [13,17] are shown here in next Figures 3 and 4. These results concern concentrations (in mg/Liter) versus time (in hours, hr) at the outlet and along the length $L = 3\text{m}$ of the tank. The concentration values are computed at selected unit points, for which experimental results are available [3]. These points were the places of observation wells at the tank entrance, at distances of $L/3$ and $2L/3$ of the tank length from the entrance, and at the tank outlet. The observation wells A, B, C, D were in the corresponding levels: $z_A = 44\text{cm}$, $z_B = 22\text{cm}$, $z_C = 1\text{cm}$ and $z_D = 44.5\text{cm}$.

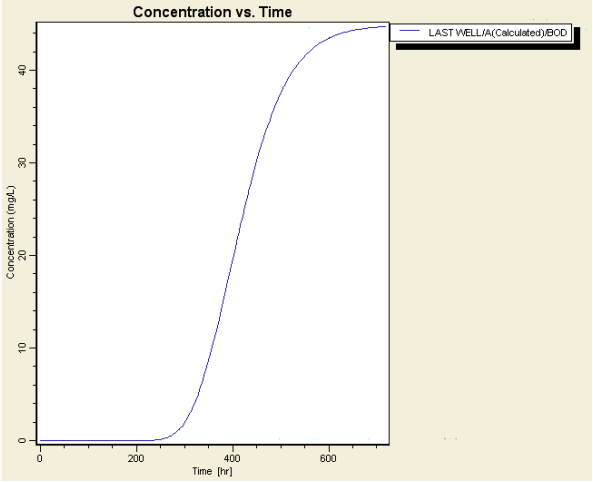


Fig. 3. Concentrations at the outlet ($x = L = 3\text{ m}$) for the Case 1

Next Table 1 summarizes some of the results for the tank MG-R: The above computed results are in good agreement with the reasoning for real wetlands. As concluded by the values in Table 1, the Case 1 is the optimal one. The same conclusion is reached by applying a Linear Programming approach, as will be presented in a next future application.

Table 1. Concentrations C_{out} (in mg/Liter) at the tank outlet

λ	0.10	0.125	0.15	0.17	0.19	0.20
Case 1	69.24	44.71	28.83	20.49	14.60	12.32
Case 2	73.53	49.54	33.59	24.93	18.63	16.13
Case 3	79.46	56.36	40.71	31.98	25.45	22.77
Case 4	71.00	46.64	30.71	22.24	16.22	13.85
Case 5	70.00	45.57	29.68	21.28	15.33	13.02

Moreover, the above computed results are in a very satisfactory agreement with the measured (experimental) ones [3]. This is shown in the following Figure 5, which concerns the Case 1 (inlet of the whole pollutant at the entrance of the tank: $x_{in} = 0.0\text{ m}$), see Akrotos [3].

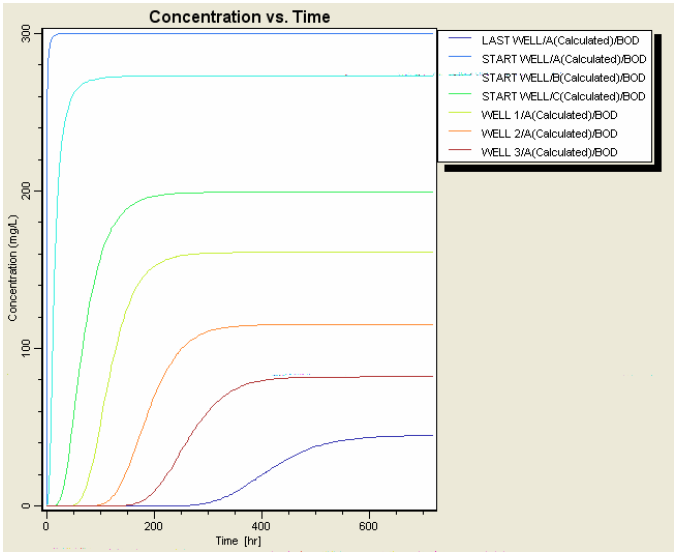


Fig. 4. Concentrations at the main points along the tank for the Case 1

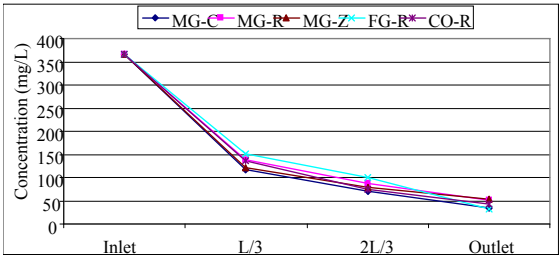


Fig. 5. Experimental average values BOD along the pilot units

4 Conclusions

A numerical simulation of flow and contaminant removal in porous media with emphasis to horizontal subsurface flow in constructed wetlands has been presented. After the mathematical modeling of the problem, the Visual MODFLOW code family, based on the finite difference method and combined with MT3DMS module, has been used for the computational investigation of rectangular pilot-scale units of horizontal subsurface flow wetlands.

The effects of the inlet positions have been numerically investigated. As the results have shown, the numerical procedure is effective for the prediction of the performance and for the optimum design of constructed wetlands.

References

1. Akratos, C.S., Tsihrintzis, V.A.: Effect of temperature, HRT, vegetation and porous media on removal efficiency of pilot - scale horizontal subsurface flow constructed wetlands. *Ecological Engineering* J. 29, 173–191 (2007)
2. Angelakis, A., Tchobanoglous, G.: *Wastewater Engineering-Natural Systems for Treatment, Reuse and disposal*. University of Crete Editions, Heraklion, Greece (1995) (in greek)
3. Akratos, C.: *Optimization of Design Parameters for subsurface Flow Constructed Wetlands by using Pilot-scale Units*. Doctoral Dissertation, Department of Environmental Engineering, Democritus University of Thrace, Xanthi, Greece (2006) (in greek)
4. Kadlec, R., Wallace, S.: *Treatment Wetlands*, 2nd edn. CRC Press, New York (2009)
5. Bear, J.: *Hydraulics of Groundwater*. McGraw-Hill, New York (1979)
6. Zheng, C., Bennett, G.D.: *Applied Contaminant Transport Modelling*, 2nd edn. Wiley, New York (2002)
7. Anderson, M.P., Woessner, W.W.: *Applied Groundwater Modeling*. Academic Press, London (2002)
8. Ewing, R.E., Lazarov, R.D.: Computational Techniques in Multiphase Flow and Transport in Porous Media. In: *Proc. Int. Conference on Mathematics and Computations, Reactor Physics, and Environmental Analysis*, Portland, Oregon, April 30 - May 4, pp. 414–420. American Nuclear Society Inc. (1995)
9. Zheng, C., Wang, P.P.: MT3DMS: A Modular Three-Dimensional Multispecies Transport Model for Simulation of Advection, Dispersion, and Chemical Reactions of Contaminants in Groundwater Systems, Documentation and User's Guide. University of Alabama, US Army Corps of Engineers Engineer Research and Development Center, Contract Report SERDP-99-1 (1999)
10. Visual MODFLOW v.4.2: User's Manual. Waterloo Hydrogeologic Inc. (2006)
11. Bear, J., Verruijt, A.: *Modeling Groundwater Flow and Pollution*. D. Reider, Boston (1987)
12. Sidiropoulou, M.: *Simulation of Flow and chemical Reactions in Porous Media with Emphasis in Horizontal Subsurface Flow Constructed Wetlands*. Master Degree Thesis, Civil Enging Dept., Democritus University of Thrace, Xanthi, Greece (2007) (in greek)
13. Liolios, K.: *Simulation of Flow and Performance Factors for Contaminant Removal in Horizontal Subsurface Flow Constructed Wetlands*. Master Thesis, Departments of Civil and Environmental Engineering, Democritus University of Thrace, Xanthi, Greece (2008) (in greek)
14. Sun, N.-Z.: *Inverse Problems in Ggroundwater Modeling*. Springer, Berlin (1994)
15. Batu, V.: *Applied Flow and Solute Transport Modelling in Aquifers*. Taylor & Francis, CRC Press, New York (2006)
16. Spitz, K., Moreno, J.: *A Practical Guide to Groundwater and Solute Transport Modelin*. Wiley, New York (1996)
17. Liolios, K., Moutsopoulos, K., Tsihrintzis, V., Akratos, C.: A computational investigation of flow and contaminant removal in horizontal wetlands. In: *Proceedings of the 6th International Congress on Computational Mechanics (GRACM)*, Thessaloniki, June 19-21 (2008)

Using Satellite Observations for Air Quality Assessment with an Inverse Model System

Achim Strunk^{1,3}, Hendrik Elbern^{1,2}, and Adolf Ebel¹

¹ Rhenish Institute for Environmental Research, Cologne, Germany

² Forschungszentrum Jülich, Jülich, Germany

³ Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

`strunk@knmi.nl`

`http://www.knmi.nl/~strunk`

Abstract. The EURAD-IM chemistry transport model and its 4d-var inverse model extension is applied to one summer and one winter episode, in order to identify the benefit of tropospheric NO₂ column retrievals for estimating near-surface nitrogen dioxide concentrations. Initial values and emission rates are jointly optimised by assimilating tropospheric NO₂ data from the OMI, while European ground based observations are used for impact evaluation. Results show a moderate improvement of surface level nitrogen dioxide estimates during the summer episode, successfully sustained after the assimilation period through emission adjustments. In the winter case, the OMI data is of limited value due to lower boundary layer heights and thus smaller impact on emission rates.

1 Introduction

The operational prediction of air quality as a computationally demanding task has become an important part of the assessment of risks for our environment and health. National and international agencies and policy makers are using numerical model applications to support their activities for air pollution reduction, for decision making and for informing the general public. However, assessing air pollution levels from global to local scale still remains a challenging mission, involving parameters with large uncertainties, e.g. emissions.

As an independent source of information, observations are playing a very important role and can therefore be used to ameliorate the simulation results. This can only be achieved in a satisfying way by applying advanced data assimilation techniques, producing consistent chemical estimates on regular grids. A strongly growing amount of environmental data is available as satellite based measurements, which are widely used for both model evaluation and data assimilation purposes by the air quality modeling community.

Through recent regulations, a stronger focus has been placed on nitrogen dioxide (NO₂) by the European Commission. Fulfilling decreasing threshold values and exceedance limits is likely to cause problems for many European cities in the coming years. At the same time, NO₂ is one of the constituents which is already available as tropospheric column retrievals since longer times. In order to estimate the

value of state-of-the-art satellite observations for air quality related assessment of nitrogen dioxide, this study assimilates tropospheric columns from the Ozone Monitoring Instrument (OMI, [1]) into the European Air Pollution Dispersion - Inverse Model system (EURAD-IM). This model system allows for jointly optimising emission rates and initial values by the four-dimensional variational method (4d-var). In the case of reactive, short-lived constituents like nitrogen oxides, optimised emission rates provide for a suitable tool to sustain valuable information in the model system, while the influence of initial states vanishes rather quickly.

This paper continues with an introduction to the model system, the emissions in use and the observational data. Section 3 shortly introduces the two case study episodes and the assimilation setup, while the results are given in Section 4. The paper is concluded by a short summary.

2 EURAD Inverse Model System (EURAD-IM)

In this study the MM5 ([2]) Version 3.7.4 has been applied to produce meteorological fields driving the chemistry transport simulations. The configuration involves the Betts-Miller cumulus scheme, the Pleim-Chang boundary layer parametrisation, the Schultz microphysics moisture and the RRTM longwave radiation schemes. The soil has been modelled using the Pleim-Xiu land surface model. Operational analyses of the ECMWF Integrated Forecast System served as boundary conditions and initial fields.

The forward CTM of EURAD-IM ([3], [4]) was applied in a non-hydrostatic configuration, employing the RACM-MIM chemistry mechanism ([5]) with 256 reactions of about 110 constituents. The number of vertical layers is 23 with about 15 layers in the boundary layer, the centre of the lowest layer being at a height of about 19 m. A rigid lid defines the boundary condition at the model top of 100 hPa. The horizontal grid spacing is 15 km, meeting the resolution of the OMI observations. The deposition scheme devised by [6] has been employed, while biogenic emissions are calculated following [7].

The data assimilation system includes the associated adjoint operators for transport and diffusion schemes ([8]) as well as for the implemented gas phase mechanisms ([9]). The adjoint (backward) version of the forward operators have been coded following [10], using the Adjoint Model Compiler ([11]) and the Kinetic PreProcessor KPP ([12]), the backward routines thus substituting their forward counterparts in the backward model part.

The adjoint modules are encoded such that both the model state parameters and emission rates are active variables, enabling to optimise both parameter sets separately or jointly, being one key feature of EURAD-IM. In order to limit the degrees of freedom, time invariant emission factors are optimised instead of hourly emission rates. These emission factors are then scaling the total emitted amount while preserving the daily cycle.

The definition of the 4d-var cost function and the control parameters, the implementation characteristics as well as the specification of the error covariance matrices are following the setup in [9]. The description of which is omitted here.

Table 1. The time frames of the two case studies during summer 2006 and winter 2007/8 were setup in the same way. The respective periods for the 3d-var (spin-up), the 4d-var (assimilation), the evaluation and the control run period are given.

	spin-up	assimilation	evaluation	control
2006	Jun 25 – 30	Jul 01 – 15	Jul 16 – 31	Jul 01 – 31
2007/2008	Dec 05 – 10	Dec 11 – 25	Dec 26 – Jan 10	Dec 11 – Jan 10

2.1 Emissions and Observations

Emission estimates have been provided using the high resolution database for Europe produced by TNO for GEMS ([13]). This database consists of gridded annual totals for NO_x , SO_2 , NMVOC, NH_3 and CO with a resolution of $1/8^\circ \times 1/16^\circ$ latitude - longitude for different polluter groups, valid for the year 2003. This inventory has been extended for this study by the ship emissions coming from EMEP inventories and being valid for the year 2000 ([14]).

The observational database comprises tropospheric nitrogen dioxide retrievals from the OMI instrument onboard the AURA satellite as well as dense routinely operated network data, gathered by the European Environment Agency and providing hourly ground-based observations of O_3 , NO_2 , CO, SO_2 , NO, C_6H_6 and NO_x . Satellite data sets were collected from the portal www.temis.nl conducted by the Royal Netherlands Meteorological Institute (KNMI). The data includes additional averaging kernel (AK) information and has been prepared using the DOMINO algorithm version 1.0.2 ([15]). OMI has a nominal pixel size of $13 \times 24 \text{ km}^2$ and an overpass time of 13:30 local time. Different evaluation studies exist for OMI data, emphasising the existence of variable bias in the OMI retrievals ([16]). These bias issues together with known problems of inferring surface nitrogen dioxide concentrations, especially during summer, will be disregarded in this paper but will be subject of further studies.

3 Episode Description and Assimilation Setup

Two case studies were selected, one to represent summerly fair weather conditions with active photochemistry, while in the second case winterly conditions with elevated surface NO_2 concentrations were chosen. For both studies of more than four weeks length, the first two weeks (15 days) were taken as analysis period with daily data assimilation, while the entailing forecast period includes the third and fourth week (16 days in total). The exact time periods are summarised in Table 1. The two episodes own common aspects in design and data flow: For provision of a balanced and accurate estimate of the initial state of the atmosphere to the 4d-var assimilation, a 6-day spin-up period has been conducted with 3d-var data assimilation 6 times per day (at 02, 06, 10, 14, 18, 22 UTC).

The assimilation interval of the 4d-var algorithm has been set to last from 09:00 UTC to 15:00 UTC, in order to encompass local overpass times of OMI

Table 2. OMI error statistics in terms of bias and rms-error ($[10^{15} \text{ molec/cm}^2]$) against model results during the two case study episodes. CNT: Control run without assimilation. BGR: Background simulation. ANA: Analysis run.

	Summer episode					Winter episode				
	Assimilation			Evaluation		Assimilation			Evaluation	
	CNT	BGR	ANA	CNT	ANA	CNT	BGR	ANA	CNT	ANA
BIAS	0.75	0.71	0.39	0.76	0.74	0.85	0.65	0.38	0.08	-0.05
RMSE	1.67	1.51	1.22	1.68	1.56	5.32	5.45	4.83	3.19	3.49

in the model domain. The analysis obtained served as initial state for an ensuing forecast until 09:00 UTC the following day, applying the analysed emission factors. Subsequently, this forecast result as well as the current emission factors are used as background state variables in the next 4d-var experiment.

The assimilation period lasts for 15 days. The following 16 days (evaluation period) is simulated without further assimilating OMI data but still comparing with observations for inferring any changes in mid-term forecast skill.

To be able to assess the overall improvements by the data assimilation experiments, an additional control run (without any data assimilation) is conducted for the whole period of 31 days, starting with the result of the spin-up simulations.

4 Results

Tropospheric nitrogen dioxide column retrievals from OMI have been assimilated throughout both case study periods. Table 2 summarises the performance of the OMI assimilation for the summer and the winter episode. There is a strong improvement of simulation performance by the assimilation of OMI observations within the assimilation period. For the summer episode, the BIAS is reduced from 0.75 (CNT) to $0.39 \cdot 10^{15} \text{ molec/cm}^2$ (ANA), while the overall RMSE can be reduced by 27 %. In contrast, the background simulation values (BGR) do not show this significant improvement, which is also the case for the adjacent evaluation forecast period. This means, that optimised emission rates slightly improve OMI model equivalents, while the main gain in forecast skill during the assimilation period is due to changes in the initial values of nitrogen oxides and those constituents which chemically influence nitrogen dioxide concentrations.

For the winter period, there is similar BIAS reduction by the analysis, but with significantly higher RMS errors obtained. During the adjacent evaluation period, there is on average almost no bias between the OMI observations and the model simulation, while the difference between the control run and the analysis based run are similar to the summer case.

The emission factors obtained by the assimilation procedure for NO_2 are shown in Figure 1. There is a clear difference in the results for the summer and the winter episode. In the summer case, there is amplification of nitrogen

dioxide emissions (grey to black) in large areas over Europe including shipping routes, cities and congested areas in Germany, Benelux, Poland, Turkey, etc., while reduction of emissions (grey to white) is linked to some high populated areas (e.g., Paris, London, Milan, Moscow) and some shipping routes in the Mediterranean. In the winter case, the amplitude of changes is much smaller, and almost only emission reduction occurs.

This distinct assimilation result could be the effect of significantly smaller observation minus background departures ($O - B$) in the winter case, which would lead to smaller changes in state variables by the assimilation. However, the average absolute departures $|\overline{O - B}|$ are $0.97 \cdot 10^{15}$ (summer) and $2.57 \cdot 10^{15}$ (winter) molec/cm², while the mean departures $\overline{O - B}$ are given by $0.71 \cdot 10^{15}$ (summer) and $0.90 \cdot 10^{15}$ (winter) molec/cm², respectively. The departure statistics does therefore not support this explanation for the smaller emission factor changes during the winter episode.

In this study exploitation of NO₂ column information rests on the vertical integration with AK weighting. Typical profiles of AK have a maximum in the upper troposphere, generally far above the boundary layer inversion, with minimum values at the surface. On the other hand, emission controlled NO₂ concentrations are highest on the lowest levels in areas subject to anthropogenic influence. In Figure 2 the AKs for the winter and summer cases are shown together with averaged mixing layer heights at noon. The fading sensitivity of the AKs towards the ground and the very low PBL height during winter thus leads to a stronger impact of the 4d-var algorithm on initial value optimisation and it strongly hampers the assimilation based optimisation of surface emission strengths.

Finally, the effect of the analyses on the surface concentrations of nitrogen dioxide is shown in Figure 3 for the summer and winter cases. The analysis is here restricted to suburban and rural stations, being most representative for the horizontal grid resolution of 15×15 km² and constituting a subset of more than

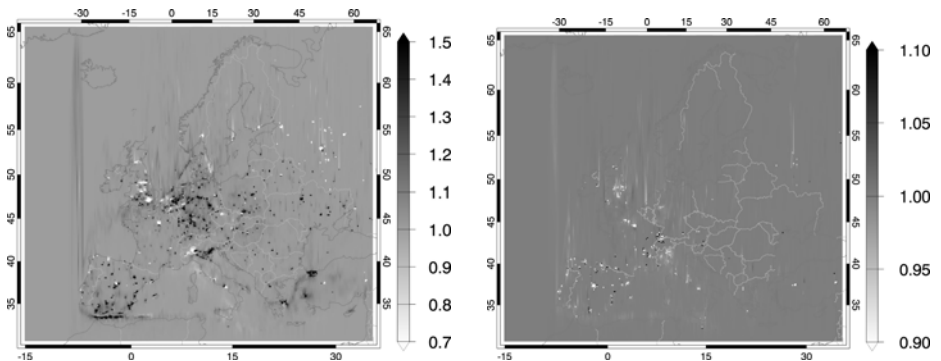


Fig. 1. Comparison of nitrogen dioxid emission factor results for the summer case study (left panel) and the winter episode (right panel). Please note the different ranges on the color scale.

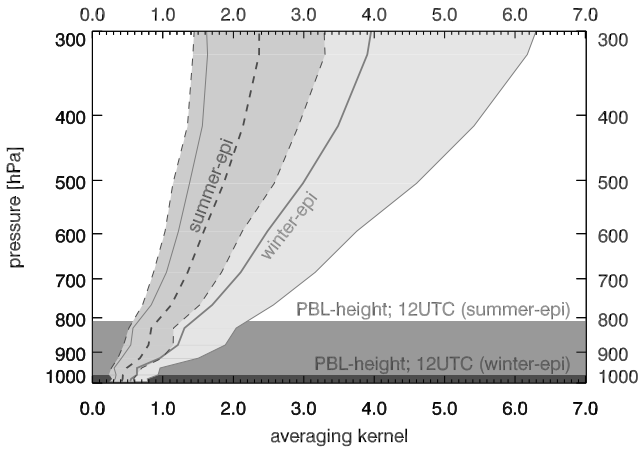


Fig. 2. Averaged AK values of the OMI observations for the summer (dashed lines) and the winter period (solid lines). Square root of variances are given as thin lines. The two horizontal, grey shaded areas show the height of the planetary boundary layer (PBL) during the winter and summer episode, averaged at 12 UTC over land covered grid cells and over the 4d-var assimilation period.

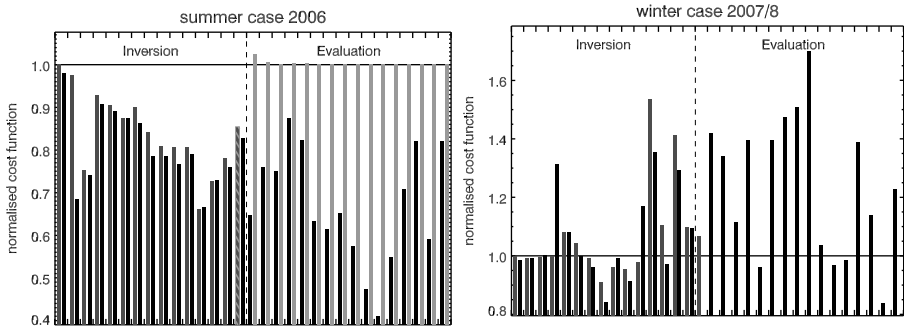


Fig. 3. Cost function values of the background simulation (dark grey) and analysis results (black), normalised by the cost function value of the control simulation without data assimilation. Values lower than 1.0 thus represent improvement of simulation skill. An additional simulation resetting emission factors has been conducted for the evaluation period (light grey) of the summer case, showing a fast relaxation of the analysis towards the control run.

one half of the total number of stations. The figure gives cost function value changes, being an objective evaluation method using the error estimates of the observations, being disregarded by bias and rms-error.

During the summer episode, the simulation skill moderately improves, both for the background simulation as well as for the analysis runs. This positive impact on the simulation performance is also sustained after the assimilation

period. The cost function values for the evaluation period can be reduced on average by about 30 %. The benefit of the OMI satellite retrievals for surface NO₂ concentrations at this subset of stations and during this evaluation period can also be seen in an rms-error reduction from 9.5 to 7.9 ppbV, which is an improvement of 17 %. In order to test whether this can be attributed to emission adjustments, an additional simulation has been conducted, resetting the emission rates to the background values after the summer assimilation period. The results are also given in Figure 3, showing a fast relaxation of this simulation to the control run (given by the value 1.0).

As expected, the winter case does not show this beneficial impact of OMI tropospheric columns for the surface NO₂ concentrations. There is no unambiguous behaviour during the assimilation period, while the evaluation period shows a clear degradation of simulation skill.

5 Summary

The assimilation of OMI tropospheric column retrievals into the state-of-the-art inverse model system EURAD-IM shows different values of the observations for assessing typical summer and winter conditions of nitrogen oxides. During summer episodes showing high PBL levels during the day, valuable information about emission rates can be inferred, allowing to preserve a positive impact of the assimilation experiment on the ensuing forecast of nitrogen oxides, especially for suburban and rural stations.

During winter conditions with low PBL levels, assimilation gain is more linked to initial values, the benefit of which vanishes after a few days. No clear and distinct characteristic of the impact of OMI NO₂ columns on the surface NO₂ concentrations can be encountered in this case.

Further investigations will focus on more surface constituents (e.g., nitrogen oxide and ozone) and will try to cope with bias issues in both observational data sets.

Acknowledgements

We acknowledge the free use of tropospheric NO₂ column data from the OMI sensor from www.temis.nl (courtesy of KNMI). Computations have been conducted at Jülich Supercomputing Centre, Forschungszentrum Jülich. This work was funded by the European Space Agency through the PROMOTE project.

References

1. Levelt, P.F., van den Oord, G.H.J., Dobber, M.R., Malkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J.O.V., Saari, H.: The ozone monitoring instrument. *IEEE Trans. Geosci. Remote Sensing* 44(5), 1093–1101 (2006)
2. Grell, G.A., Dudhia, J., Stauffer, D.R.: A description of the fifth-generation Penn State/NCAR mesoscale model MM5. Technical report, National Center for Atmospheric Research (1994)

3. Hass, H.: Description of the EURAD Chemistry-Transport-Model version2 (CTM2), vol. 83. Mitteilungen aus dem Institut für Geophysik und Meteorologie der Universität zu Köln (1991)
4. Memmesheimer, M., Friese, E., Ebel, A., Jakobs, H.J., Feldmann, H., Kessler, C., Piekorz, G.: Long-term simulations of particulate matter in Europe on different scales using sequential nesting of a regional model. *Int. J. Environm. and Pollution* 22(1-2), 108–132 (2004)
5. Geiger, H., Barnes, I., Bejan, I., Benter, T., Spittler, M.: The tropospheric degradation of isoprene: an updated module for the regional atmospheric chemistry mechanism. *Atmos. Environ.* 37(11), 1503–1519 (2003)
6. Zhang, L., Brook, J.R., Vet, R.: A revised parameterization for gaseous dry deposition in air-quality models. *Atmos. Chem. Phys.* 3, 2067–2082 (2003)
7. Guenther, A.B., Zimmerman, P.R., Harley, P.C., Monson, R.K., Fall, R.: Isoprene and monoterpene emission rate variability: Model evaluations and sensitivity analyses. *J. Geophys. Res.* 98(D7), 12609–12617 (1993)
8. Elbern, H., Schmidt, H., Ebel, A.: Variational data assimilation for tropospheric chemistry modeling. *J. Geophys. Res.* 102(D13), 15967–15985 (1997)
9. Elbern, H., Strunk, A., Schmidt, H., Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.* 7, 3749–3769 (2007)
10. Talagrand, O.: The use of adjoint equations in numerical modelling of the atmospheric circulation. In: Griewand, A., Corliss, G.G. (eds.) *Proceedings of Workshop on Automatic Differentiation of Algorithms: Theory, Implementation and Application* (1991)
11. Giering, R.: Tangent linear and Adjoint Model Compiler, Users manual TAMC Version 5.2 (September 1999)
12. Sandu, A., Sander, R.: Technical note: Simulating chemical systems in Fortran90 and Matlab with the Kinetic PreProcessor KPP-2.1. *Atmos. Chem. Phys.* 6, 187–195 (2006)
13. Visschedijk, A., Zandveld, P., van der Gon, H.D.: A high resolution gridded European emission database for EU integrated project GEMS. TNO report 2007-A-R0233/B, TNO (March 2007)
14. Vestreng, V.: Review and revision. Emission data reported to CLRTAP. Technical report, EMEP MSC-W, Norwegian Meteorological Institute, Oslo, Norway (2003)
15. Boersma, K.F., Eskes, H.J., Veefkind, J.P., Brinksma, E.J., van der A, R.J., Sneep, M., van den Oord, G.H.J., Levelt, P.F., Stammes, P., Gleason, J.F., Bucsela, E.J.: Near-real time retrieval of tropospheric NO₂ from OMI. *Atmos. Chem. Phys.* 7(8), 2103–2128 (2007)
16. Lamsal, L.N., Martin, R.V., van Donkelaar, A., Celarier, E.A., Bucsela, E.J., Boersma, K.F., Dirksen, R., Luo, C., Wang, Y.: Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: Insight into the seasonal variation of nitrogen oxides at northern midlatitudes. *J. Geophys. Res.* 115(D05302) (2010)

Distributed Software System for Data Evaluation and Numerical Simulations of Atmospheric Processes

Atanas T. Terziyski and Nikolay T. Kochev

University of Plovdiv, 24, Tzar Assen Str., Plovdiv 4000, Bulgaria

Abstract. A distributed software system for numerical simulations of atmospheric physicochemical processes is presented. It is a multi-layer Java based system for theoretical investigation of complex interactions of atmospheric trace gases and ice particles. The simulations are based on the fundamental theory of Langmuir adsorption and second Fick's law applied for adsorption, desorption and diffusion processes. The system consists of three basic layers: (1) input/output interface layer, (2) dispatcher layer, (3) grid-based layer for simulations distributed over multiple machines. The core software module used in level (3) is based on previously published by us software prototype for simulations of adsorption, desorption and diffusion in a closed system and Flow Tube Reactor. The main task of the current distributed system is to derive numerical estimations of several significant constants: adsorption/desorption rates, ice entry rate, ice bulk diffusion coefficient and etc. The constants are estimated by comparison of experimental signals from a Flow Tube Reactor and simulations results from the system described in this paper. The difference between both curve profiles is minimized by an exhaustive search in a multi-dimensional parameter space which represents all possible values of the physicochemical constants. The dispatcher layer of the system defines several regions of the multi-dimensional parameter space. For each region, a separate task is configured and dispatched to a node from a computer GRID or cluster. The entire parameter space is searched in a parallel manner and after that all results are united in order to find the global minimum of the difference between experimental and simulated curves. The results are printed via the input/output software layer. Example kinetic simulations performed by the software system are presented and discussed.

Keywords: distributed software system, atmospheric chemistry, simulation, adsorption, desorption, diffusion, modelling.

1 Introduction and Experimental Setup

The deeper understanding of atmospheric processes has become an important topic for physicists and chemists in the last decades. The joint research is carried out by worldwide scientists in few major directions such as studying the heterogeneous processes occurring in different altitudes, understanding the ice and ice

structure as adsorbent and ice bulk diffusion. Both experimental and theoretical approaches have been used for deriving thermodynamical and kinetic parameters characterizing atmospheric processes. The atmospheric processes are studied developing various kinetic models of Coated Wall Flow Tube (CWFT) reactors [1,2]. Reactor experimental results contain useful information but usually it is in hidden form. Deriving the essential parameter information could be done only comparing the experimental signal with appropriate mathematical simulation of the studied process. In this paper we present version 1.0 of software system ADDESSA (ADsorption DEsorption Simulation system for Signal Analysis). ADDESSA is an integral part of on-going project for study of atmospheric physicochemical processes. The experimental studies of adsorption/desorption with diffusion/segregation kinetics and thermodynamics has been laboratory performed on CWFT reactor. The apparatus consists of a tube cooled down at low and constant temperatures in accordance with atmospheric conditions. A movable injector that ends with a tiny nozzle injects the measured trace gas along the reactor tube while it moving. The carrier gas flows laminarly toward the mass spectrometer detector. Initially ice film is generated by deposition of gaseous water upon injection of water vapour through the sliding injector at low temperatures of the reactor wall.

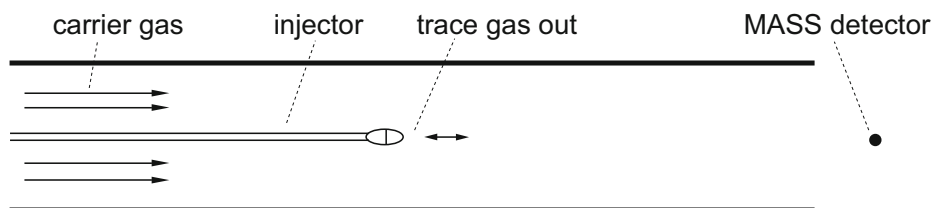


Fig. 1. Schematic representation of CWFT reactor

During this procedure the injector is slowly moved in order to generate a smooth surface film. The experiments later on are performed by moving the injector backwards and forwards with a constant speed. After each movement a certain time has to elapse during which the adsorption and diffusion equilibrium is reestablished. The profiles are the plots of detector signal against the laboratory time. The typical measurement profile consists of five consecutive stages according to the position and speed of the movable injector. The detected MASS signal is a sum of the molecules from the injector flow and the previously deposited molecules in the reactor which took part in adsorption, desorption and diffusion processes. The shape and the time scale of the signal change in each stage strongly depends on the rate coefficients, ice film thickness, ice and bulk concentration, temperature range, flow and injector speed, etc. The experimental conditions are stepwise varied in the laboratory measurements [3].

2 Mathematical Model of Atmospheric Physicochemical Processes

The relation between the gas phase and adsorbed molecules is given by the following system of ordinary differential equations (1). The term $-S/V$ represents the surface to volume ratio concerning the given geometry. The rate coefficients of adsorption and desorption respectively k_{ads} [cm^3s^{-1}] and k_{des} [s^{-1}] are given as parameters, c_g [cm^{-3}] is the gas phase, c_s [cm^{-2}] the surface and $c_{s,max}$ [cm^{-2}] the maximal surface concentrations.

$$\begin{cases} \frac{dc_s}{dt} = k_{ads}c_g(c_{s,max} - c_s) - k_{des}c_s \\ \frac{dc_g}{dt} = -\frac{S}{V}(k_{ads}c_g(c_{s,max} - c_s) - k_{des}c_s) \end{cases} \quad (1)$$

The diffusion is defined by a diffusion coefficient D [cm^2s^{-1}] which stands for a number of molecules passing through a certain area for a certain time period. The diffusion simulation is based on the second Fick's law (2):

$$\frac{\partial c_b(x,t)}{\partial t} = D \frac{\partial^2 c_b(x,t)}{\partial x^2} \quad (2)$$

where $c_b(x,t)$ [cm^{-3}] is the bulk phase concentration as function of time t and bulk depth x . The connection between gas phase and bulk phase is defined via Neumann type boundary condition for equation (2) where the derivative of $c_b(x,0)$ is proportional to c_s and $(c_{b,max} - c_b(x,0))$. The two equations are applied stepwise along the tube reactor in order to update kinetically the measured gas phase concentration by the detector [2]. Each simulation calculation is performed for particular values of a set of parameters $\{p_i\}_{i=1..s}$. This set typically includes constants that characterize the thermodynamic and kinetic properties of studied chemical compound e.g. k_{ads} , k_{des} , $c_{s,max}$, D and $c_{b,max}$. The simulation result is a function of these parameters $S_{simulated} = f(p_i)$. The parameter values are preliminary unknown and it is assumed that their best estimation minimizes the difference between simulated and measured signals RMS_{Error} .

$$\text{RMS}_{Error} = \sqrt{\sum_{i=1}^n (S_{simulated}^i - S_{MASS}^i)^2} \quad (3)$$

The influence of parameters $\{p_i\}$ over RMS_{Error} can not be expressed in analytical form. The automatic determination of the optimal values of $\{p_i\}$ is performed by an exhaustive systematic search scanning each point from a multidimensional grid (Fig. 2). Simulations are performed for each point from this grid. The search for minimal RMS_{Error} requires huge computational resources especially when numerical stability and accuracy concerns are handled. Quality simulations demand small values of the differentiation steps (dx and dt) for the numerical solving of the differential equations (1) and (2) and large number of reactor segments and ice bulk layers. The search process can be parallelized by

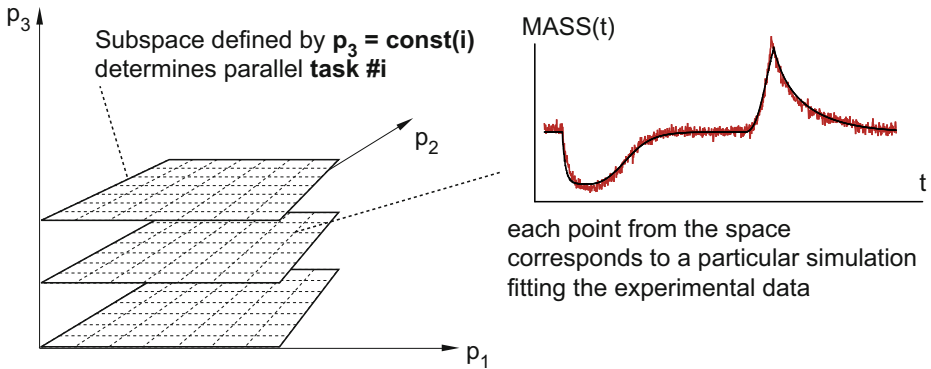


Fig. 2. Systematic searching of the parameter space

dividing the parameters space of several regions and performing simulations for each space region on a different machine.

3 Software Architecture

The described above numerical solutions are implemented by means of distributed multilayer software architecture (Fig. 3). The system includes three basic layers: (1) Input/Output Layer, (2) Dispatcher Layer and (3) Computation Layer. Most of the software modules are implemented in Java language. Some most critical modules performing the core simulations are additionally implemented in C++. Layer (1) implements the interface with the system users. Simulations can be configured and started via a web interface or standalone application with graphical user interface (GUI). If particular user has an account for the front-end node from the Dispatcher Layer, the job can be started directly by putting a configuration file into the Master Dispatcher input folder.

Dispatcher Layer (2) practically consists of two sub layers called Master Dispatcher (Fig. 4) and ADDESSA Dispatcher. The main function of the Master Dispatcher is to perform the basic interface to a computer cluster or GRID architecture. The web users of the ADDESSA system do not have direct access to the computer cluster. Layer (2) provides a transparent work for the user with the ADDESSA system without involving the user into the details of the computational layer. Master Dispatcher is a configurable module implemented in Java which can perform a different types of jobs defined in a configuration file. It is a multi-thread server application where separate threads handle the input jobs, the job results (outputs) and helper utility tasks. The input to the Master Dispatcher can be done in three ways: getting the job file from a local directory, listening for a job file in directory from a remote machine or inputting job via TCP port listener. Each job is characterized by its type which determines the preparation of the job specific input and executed commands for this type of job. After a job is started the output listener waits for a specific output.

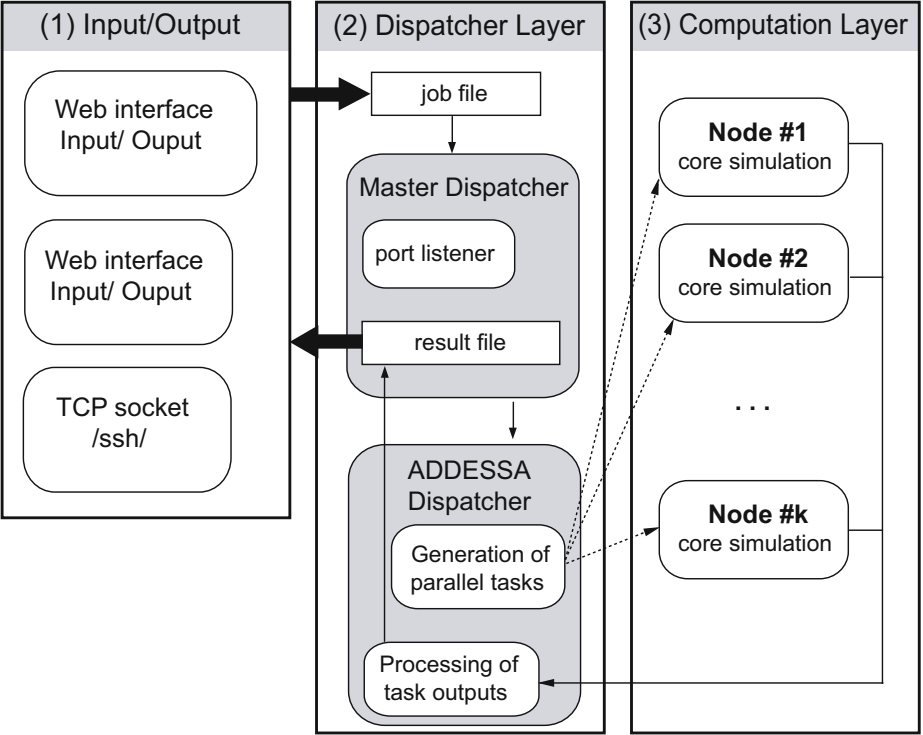


Fig. 3. Multi-layer architecture of ADDESSA software system

The output listener works analogously to the input listener. The job outputs can be processed in several ways: output is saved, output is send to a remote machine, external application is started etc. Master Dispatcher can be configured to send the original input file to a remote machine. In this way Master Dispatcher can be used as a job router.

ADDESSA Dispatcher is a server application implemented on top of Master Dispatcher. It can perform all configurable common purpose jobs as Master Dispatcher does plus several specific functions which handle the simulation calculations. ADDESSA Dispatcher uses functionality from the ADDESSA core library in order to generate a set of parallel tasks and to combine all outputs from these tasks (Fig. 5a). For each parallel task ADDESSA Dispatcher starts a separate instance of a console application which searches a region from the parameter space for an optimal simulation data fit (Fig. 2). These applications are based on the ADDESSA core library as well. According to the computer cluster hardware and software implementation all application instances are distributed to run on different computer nodes with approximately equal free CPU power. The latter guaranties a relatively small delay between the fastest and slowest parallel tasks. ADDESSA Dispatcher has a specialized Output Listener for the parallel tasks where each task application is expected to return a result

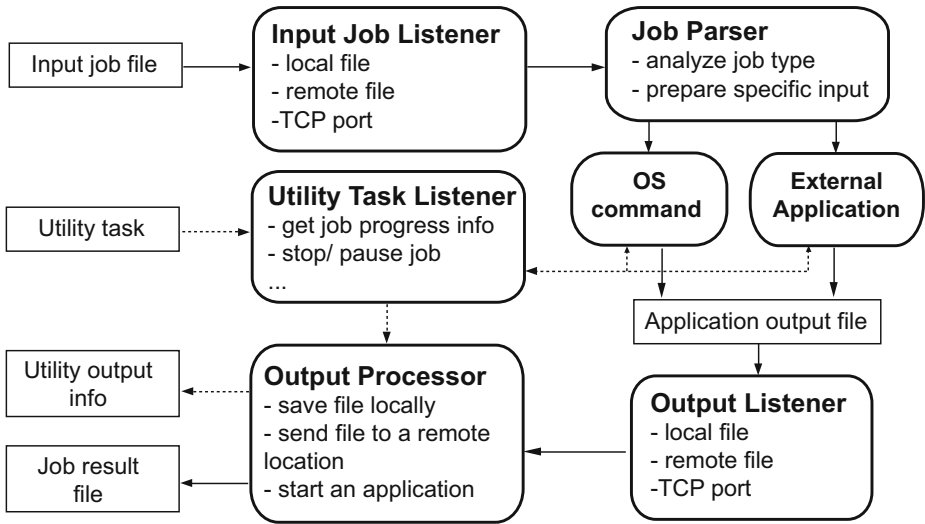


Fig. 4. Master Dispatcher Software Module

file with specific name and location. When all outputs are gathered the Output processor is calculating the best parameters for the experimental data fit by finding the minimal RMS Error from all regions. The result is sent back to the Master Dispatcher and then back to the user interface application from Layer (1). ADDESSA core simulation library consist of several packages which implement the basic functionality: simulation of adsorption, desorption and diffusion; visualization of the result simulations; interactive adjusting of the simulations parameters via GUI; automatic search for optimal parameters; parallelization of the search process.

The library is implemented complying Object-Oriented Programming where each main component is represented by one or more classes. The main input to the program is given by a configuration file where the basic parameters of the experimental process are specified as well as the simulation parameters. The main application calls a module which represents CWFT reactor. The reactor contains an array of objects of the type Reactor Segment. Reactor segment is used to store the state of a particular part of the reactor. Additionally when diffusion is applied, Reactor Segment module solves numerically the equations of the Second Fick's law. Each Reactor Segment contains a number of Bulk Layers. Reactor Manager is a GUI (Graphical User Interface) module which can be used to control the simulation process for a single simulation. All parameters inputted from the configuration file can be varied in this module. This GUI application can be used for a preliminary rough approximation of simulation parameters which can help to chose more precise intervals for the automatic parameter search.

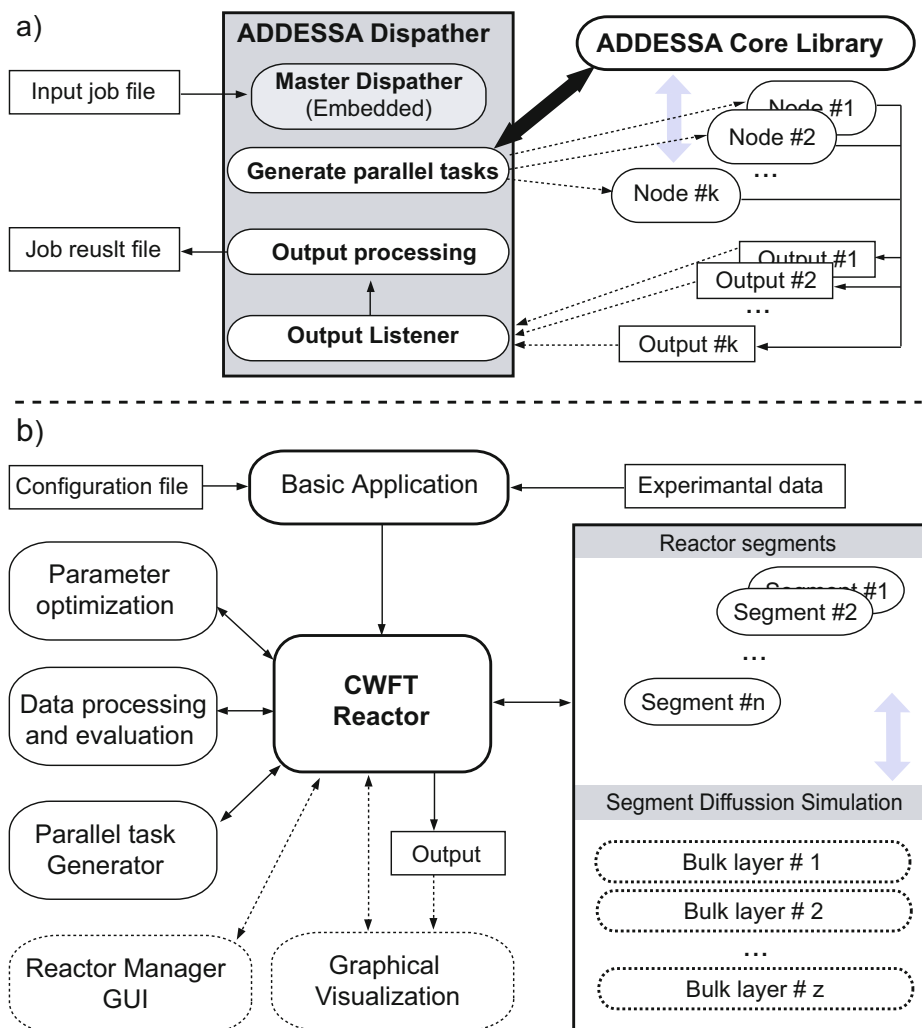


Fig. 5. Flow chart of a) ADDESSA Dispatcher, and b) ADDESSA core simulation library

4 Software System Application

In this paper we present an application of ADDESSA system on top of MADARA cluster. MADARA stands for "Modelling in ADvAnced Research Actions" and it is a joint research national project aimed to establish a Computing Centre with an up-to-date computer system and advanced software for modeling and simulations in the field of chemistry and materials sciences. It supports scientists in areas such as molecular design, new materials and nanotechnology, modeling and large-scale simulations of complex chemical systems etc. Currently MADARA is

a rack optimized cluster of 54 dual socket rack servers PRIMERGY RX200 S5. Master Dispatcher along with the Web server is deployed at the University of Plovdiv (Fig. 3). ADDESSA Dispatcher server is deployed at the MADARA front end node. ADDESSA Dispatcher uses MADARA dispatcher system to run the generated parallel task on appropriate MADARA nodes. As an example of the application, the simulated spectrum on the right side of figure 2 represents a laboratory measurement of acetone adsorption on ice surfaces under atmospheric conditions. ADDESSA system has been applied to fit the raw experimental data. A stepwise optimization has been done by varying the parameters as adsorption, desorption rates and maximal surface coverage. The corresponding rate coefficients were estimated for each measurement. This procedure was further more applied to the experimental data taken at different temperatures, concentration, ice properties, etc. in order to obtain the thermodynamical data. Both kinetic and thermodynamical values estimated by this distributed model have a very good agreement with other data reported in the literature [5].

Acknowledgements

We would like to thank the Bulgarian National Fund for Scientific Research NFNI (project MU02/12) for supporting this study. This paper is also supported by NATO reintegration grand PDD(CP)-(CBP.EAP.RIG 982653).

References

1. Behr, P., Terziyski, A., Zellner, R.: Reversible Gas Adsorption in Coated Wall Flow Tube Reactors. Model Simulations for Langmuir Kinetics. *Z. Phys. Chem.* 218, 1307–1327 (2004)
2. Terziyski, A., Kochev, N.: Modelling Of Surface And Bulk Processes In The Atmosphere. *Journal of International Scientific Publications, Issue Ecology and Safety* 2, Part 1, 341–357 (2008)
3. Behr, P., Terziyski, A., Zellner, R.: Acetone Adsorption on Ice Surfaces in the Temperature Range $T = 190\text{--}220\text{ K}$: Evidence for Aging Effects Due to Crystallographic Changes of the Adsorption Sites. *J. Phys. Chem. A* 110(26), 8098–8107 (2006)
4. Modelling in ADvAnced Research Actions. MADARA, <http://madara.orgchm.bas.bg>
5. Somnitz, H.: Quantum chemical studies of the adsorption of single acetone molecules on hexagonal ice Ih and cubic ice Ic. *Phys. Chem. Chem. Phys.* 11, 1033–1042 (2009)

Advanced Numerical Tools Applied to Geo-environmental Engineering - Soils Contaminated by Petroleum Hydrocarbons, a Case Study

Maria Cristina Vila, J.M. Soeiro de Carvalho, and António Fiúza

University of Porto, Faculty of Engineering, CIGAR
Rua Roberto Frias, 4200-465 Porto, Portugal

Abstract. Contaminated soils can be considered as a heterogeneous, anisotropic and discontinuous geo-system, whose properties vary in time and space. Focusing on the remediation of a real contaminated site (a refinery located in northern Portugal), soil samples contaminated with petroleum hydrocarbons were subject to laboratory studies. The results of contaminant degradation kinetics tests led to the development of a distributed parameter model describing simultaneously the time evolution of biomass and contaminant degradation. Several phenomena were globally taken into account in this model: the volatilization, a fast kinetics component, a slow kinetics component and the refractory hydrocarbons for the time scale used in the experiments. To complement kinetics tests the soil contaminated was submitted to respirometric tests. The *a priori* unpredictability of the respirometric results justified the continuous measurement of oxygen and carbon dioxide concentrations and of temperature in the soil atmosphere, resulting in a huge volume of data. Several mathematical techniques were used in respirometry data treatment, namely: time series, system identification and wavelets theories.

Keywords: contaminated soils, biodegradation models, respirometry data, time series analysis, system identification models, wavelets.

1 Introduction

The development of a mathematical model describing any physical phenomena requires a good understanding of the processes involved. Usually it is based on balance equations (mass and energy balance equations). However, the complexity of the model varies according to the natural - artificial dichotomy of the objects. Anthropogenic objects (artifacts developed to satisfy certain needs of man) are associated with a substantially higher rationality than that associated with natural physical objects. Soils are natural particulate systems, which are heterogeneous and anisotropic. Those contaminated by petroleum hydrocarbons are widely spread in earth surface. Organic contaminants in soils and groundwater are often used as carbon and energy source in the metabolism of

autochthonous aerobic heterotrophic microorganisms. This work can be divided in two parts, corresponding to different scales and setup of the experimental tests:

Kinetics tests in liquid cultures. A microbial consortium was selected and enriched, and shown to be capable of degrading the contaminant. It was submitted to kinetic tests in Erlenmeyer flasks. Thus the biodegradation model was validated;

Respirometry tests in soil samples. Going from micro scale to meso scale, samples of a real contaminated soil were submitted to respirometry tests in a TR8 Sablesys respirometer. A controlled flow of air was induced in the atmosphere in the vicinity of contaminated soil and the concentrations of oxygen and carbon dioxide and temperature were registered allowing to monitor the activity of microorganisms in the contaminant biodegradation process.

In the field of contaminated soils, mathematical models developed to date are variations of the transport and fate of contaminants in porous media model. These models describe the variation of the concentration of each contaminant in a time-space domain, and may include terms such as reactive sorption, decay, and biodegradation (using Monod (1949) or Michaelis-Menten (1913) models). The Monod model is broadly validated at laboratory scale when working with pure contaminants and pure cultures of microorganisms. However, in nature things are quite different: the contamination consists almost always of a mixture of chemical species and the microbial population is also, in most cases, formed of a consortium of different strains of microorganisms. Here, the validation of Monod model becomes a very difficult task.

2 A Mathematical Model for Biodegradation

The model we propose for the biodegradation of petroleum compounds in contaminated soils is based on the assumption that the decrease in concentration of the contaminant over time is done mainly through two routes [1], namely: (a) abiotic decay and (b) biodegradation. The total substrate concentration at each time step is equal to the sum of two parts, one corresponding to the biodegradable substrate and the other corresponding to non-biodegradable substrate. In time, the biodegradable portion will be zero and the non-biodegradable portion will have an asymptote. The abiotic decay occurs at two different rates: volatile compounds with low solubility have a higher degradation rate; the semi-volatile compounds are more soluble and degrade slower. The corresponding mathematical expressions are:

$$\frac{dM_t}{dt} = k_1 M_t S_{bi} - k_2 M_t^2 \quad (1)$$

$$\frac{dS_{bi}}{dt} = -\frac{1}{Y} k_1 M_t S_{bi} \quad (2)$$

$$\frac{dS_{ab}}{dt} = -k_3 S_{ab} + k_4 \quad (3)$$

$$S = S_{bi} + S_{ab} \quad (4)$$

being M_t the concentration of microorganisms [ML^{-3}], S the total concentration of substrate [ML^{-3}], S_{bi} the biodegradable portion of substrate [ML^{-3}], S_{ab} the volatile portion of the substrate [ML^{-3}], Y is the biological yield, k_1 [$M^{-1}L^{-3}T^{-1}$], k_2 [$M^{-1}L^{-3}T^{-1}$] and k_3 [T^{-1}] are kinetics parameters, and k_4 [$ML^{-3}T^{-1}$] is a constant. Equations (1) and (2) are related to the contaminant biodegradation. They model the interaction between the biodegradable petroleum substrate and the biomass. The biomass follows a modified Monod kinetics: the degradation rate is proportional to the product of the biomass M_t by the substrate concentration S_2 , complemented by a logistic term: the rate decreases proportionally to the square of the biomass concentration.

It is possible to experimentally measure the initial biomass concentration $M_t(0)$ and to estimate the approximate initial concentration for the biodegradable component $S_{bi}(0)$. The kinetic constant k_1 can be easily derived: as we deal with a second order kinetics the initial half-life is inversely proportional to the initial concentration.

$$T_{1/2} = \frac{1}{k_1 S_{bi}(0)} \Rightarrow k_1 = \frac{1}{T_{1/2} S_{bi}(0)} \quad (5)$$

For the determination of parameters k_3 and k_4 relating to the behavior of non-biodegradable fraction of substrate, beginning with integrating analytically equation (3) for the known initial condition $S_{ab}(0)$, which results in,

$$s_{ab}(t) = S_{ab}(0)e^{-k_3 t} + \frac{k_4}{k_3} (1 - e^{-k_3 t}) \quad (6)$$

When time tends to infinity, S_{ab} tends to an asymptotic concentration $a = \frac{k_4}{k_3}$. A linear regression between $\ln \left(\frac{S_{ab}(0)-a}{S_{ab}(t)-a} \right)$ and time (t) determines the parameter k_3 .

2.1 Results: Fitting Experimental Data

A system of three ordinary differential equations (1),(2) and (3) was integrated numerically using the fourth order Runge-Kutta method. Model exploitation showed that it is highly sensitive to the initial biomass concentration. Figure 1 illustrates the results showing biodegradable, volatile, total substrate and biomass components after testing different time steps and parameters values. Intrinsic properties of the contaminant (crude oil) make difficult the biomass quantification in liquid medium. However, experimental determinations of substrate concentration (diesel and crude oil) are in accordance with the simulated results.

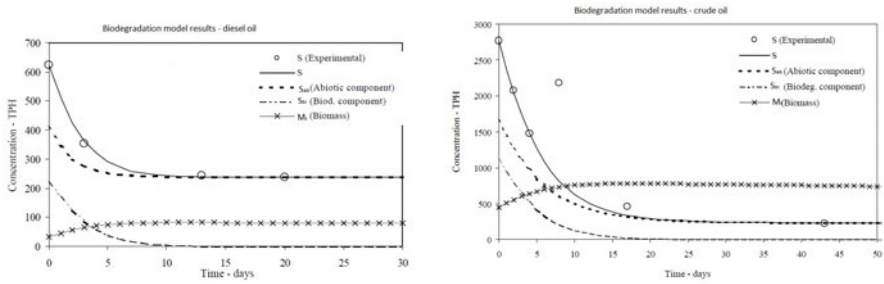


Fig. 1. Biodegradation model results for soil contaminated by diesel and crude oil respectively

3 Respirometry Data Treatment

Extensive data concerning the oxygen and carbon dioxide concentrations in the atmospheric vicinity of the soil allows an easy quantification of some of the essential parameters of the process: a detailed determination of its kinetics, a complete time differentiation of all the evolution phases involved (adaptation, active degradation and closing stage), the quantification of the cumulative usage of oxygen in the overall process and in each phase, the calculation of the stoichiometry of the biodegradation process through the ratio between CO_2 production and O_2 consumption. This information can still be interrelated to periodical chemical analysis of the contaminant concentration in the soil and to measurements of the time evolution of the biomass[2].

3.1 Time Series Analysis

Time evolution of oxygen concentration inside a reactor with contaminated soil clearly pointed to the existence of a daily cyclical component. This behavior is the subject of study of time series, which deals with phenomena that exhibit trend and seasonal components. Time series studies can follow two non exclusive approaches:

Time domain. The time domain approach is based on the assumption that the correlation between adjacent time points is best explained in past values.

Frequency domain. Frequency domain studies are particularly significant when data exhibits periodic sinusoidal variations. These periodic variations are almost always caused by natural physical phenomena, in our case the daily seasonality observed in the oxygen concentration measured inside the reactor is clearly a biological phenomenon.

The figures 2 to 4 show the measured signal of the main variables in respirometry of contaminated soils: oxygen and carbon concentrations and temperature, and their extracted seasonal component, revealing their cyclical behavior. Focusing on oxygen signal, figure 2-a), it is possible to identify three distinct tendencies.

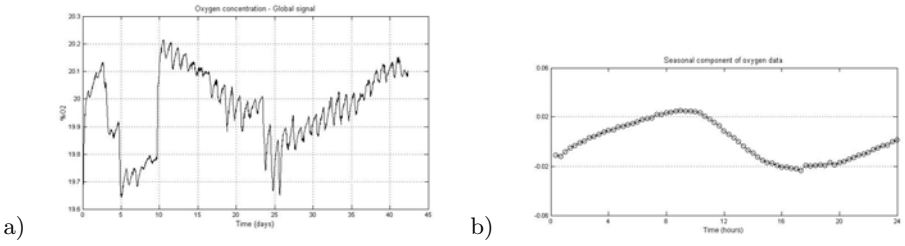


Fig. 2. Measured time evolution a) and corresponding seasonal component b) of oxygen

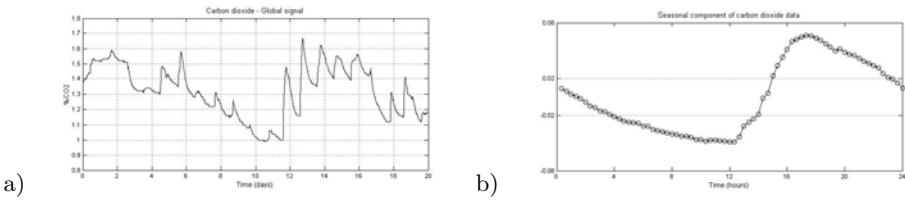


Fig. 3. Measured time evolution a) and corresponding seasonal component b) of carbon dioxide

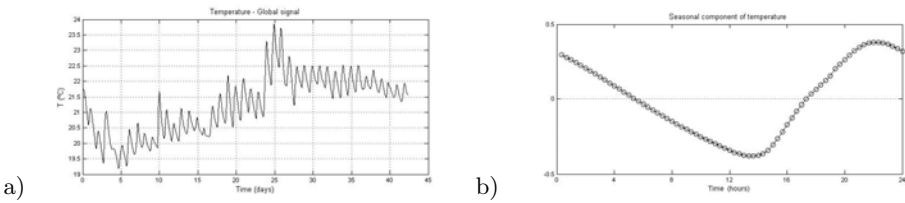


Fig. 4. Measured time evolution a) and corresponding seasonal component b) of temperature

An initial unstable period corresponding to bacterial adaptation lasting for 10 days (lag period), an active period of biodegradation lasting from the 18th to the 28th day corresponding to the exponential phase of bacterial growth and a period of decreasing biological activity from the 28th day onwards (stationary and death phases). The seasonal component of the analyzed signals will be used later in the application of the wavelets theory. The second descending phase on oxygen concentration signal (days 18 to 28 in the first graphic in figure 2-a), is more interesting due to its association to the actual soils bioremediation. It will be the base signal to model as a linear system by system identification theory.

3.2 Determination of a Model Structure Using System Identification

A Linear system produces observable signals when submitted to different interacting input variables [3]. These observable signals are called responses or

outputs. Beside the observable signals, we must consider the external stimuli, also called disturbances or inputs, that affect the system. The external stimuli constitute the manipulated variables. The purpose of System identification is to build a black box quantitative model predicting an implicit time evolution of biodegradation through the relationship between some explicit input variables (concentrations and temperature at the inlet) and the oxygen concentration at the outlet. The Matlab©System Identification Toolbox was used as the main software to test different structures from a family of transfer functions models, namely ARX, ARMAX, Output Error and Box-Jenkins. Generally, the structure is described in terms of a generalized Auto Regressive Moving Average Model (ARMA), mathematically described by:

$$y(t) = \frac{B(q)}{F(q)}u(t - nk) + \frac{C(q)}{D(q)}e(t) \quad (7)$$

where t is continuous time and k is discrete time, $y(t)$ is the output, $u(t)$ the input, $e(t)$ is a disturbance random variable, $B(q)$, $F(q)$, $C(q)$ and $D(q)$ are polynomials ordered in descending powers of the shift operator represented as q : $A(q) = 1 + a_1q^{-1} + a_2q^{-2} + \dots + a_nq^{-n}$.

The following algorithm shows the way we used the Matlab©toolbox in the stochastic modeling of respirometry results:

```

read signal
choose the section of the signal that will serve to build the model
take full signal as data for model validation
for all the available models (AR, ARMAX, Box-Jenkins) do
    choose model order : order = 2 to 8
    write the equations (according to the data and the chosen order)
    apply the optimization criterium (least squares minimization)
    validate the model
    view plot for the model outputs (real and simulated)
    view plot for residuals autocorrelation
    keep the best results
end for
```

Lets present two examples illustrating a SISO (Single Input Single Output) and a MISO (Multiple Input Single Output) model. The SISO model (figure 5) used temperature signal as input variable to predict oxygen concentration as the output variable. The MISO model (figure 6) used temperature and atmospheric oxygen concentration signals as input variables to predict oxygen concentration inside the reactor as the output variable.

3.3 Spectral Analysis: Fourier and Wavelets Transforms

Wavelets theory has its basis on the windowed Fourier transform, although the wave window is significantly different. Fourier analysis consists in transforming

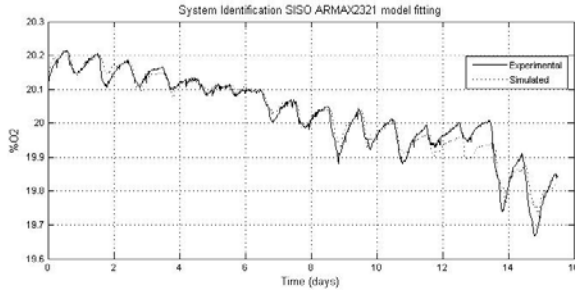


Fig. 5. Results of a ARMAX44221 model for a SISO structure

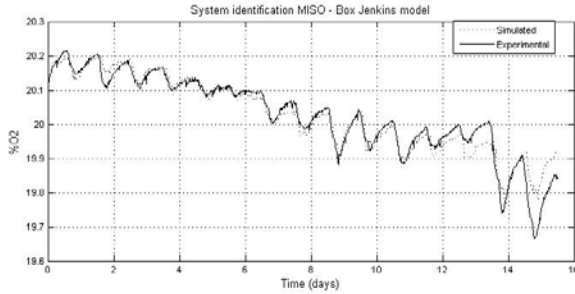


Fig. 6. Results of a BJ4222 model for a MISO structure

one signal in a set of sinusoidal waves with different wavelengths. Similarly, the wavelet analysis consists in transforming one signal in a set of scaled and translated versions of the original (or mother) wavelet. Starting from a mother-wavelet it is possible to cover all time-frequency domain through its successively dilatation and translation [4], as shown by equation (8).

$$\psi_{(u,s)}(t) = \frac{1}{\sqrt{s}} \psi \left(\frac{t-u}{s} \right) \quad (8)$$

where $\psi_{u,s}$ is the scaled (with s) and translated (with u) mother wavelet. The scaling function displays notorious advantages, especially when we are interested in details located in the high frequencies and consequently masked by white noise. In our research we applied wavelet analysis to three different situations:

- (i) As a way of detecting the main features of biodegradation through analysis of pseudo-scalograms;
- (ii) As a tool for denoising the original signals allowing them to be used in subsequent mathematical studies and simultaneously reducing the amount of data;
- (iii) As a method of detecting variation patterns at small scales.

Lets focus on the first item. As mentioned above one can notice three different events in oxygen concentration signal. These can be clearly noticed using the

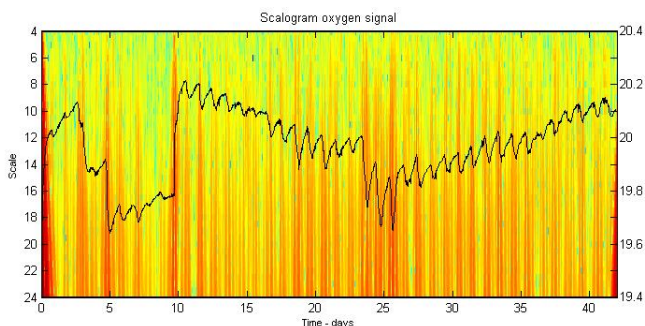


Fig. 7. Pseudo-scalogram of the signal respecting to oxygen content

wavelets through the scalogram (figure 7). Around the tenth day the scalogram evidences an energy spot spread for all the frequencies. On the 23rd day a concentration of frequencies may be seen. The daily cycle of biological activity is also evident in the pseudo-scalogram.

4 Conclusions

Modeling biologic activity in the field of contaminated soils is a promising task as autochthonous microorganisms have an important role in soil remediation. The model describing the kinetics of petroleum hydrocarbons biodegradation adequately fits the main involved variables. Respirometry data analysis using time series and spectral theory gave us the essential knowledge about the life cycle of microorganisms allowing to establish relationships between the two different studied scales in future work.

References

1. Vila, M.C., Nunes, O.P., Fiúza, A.: A model of biodegradation of crude oil in soils. In: Proceedings of the First Bioremediation Conference, Chania, Greece, pp. 1–4 (2001)
2. Vila, M.C., Fiúza, A.: An insight into soil bioremediation through respirometry. *Environment International* 31, 179–183 (2005)
3. Liung, L.: System identification theory for the user. Prentice Hall, New Jersey (1987)
4. Mallat, S.: A wavelet tour of signal processing, 2nd edn. Academic Press, London (1999)

Richardson Extrapolated Numerical Methods for Treatment of One-Dimensional Advection Equations

Zahari Zlatev¹, Ivan Dimov², István Faragó³, Krassimir Georgiev²,
Ágnes Havasi⁴, and Tzvetan Ostromsky²

¹ Nat. Environmental Research Institute, Aarhus Univ., Roskilde, Denmark

² Institute of Information and Communication Technologies, Bulgarian Acad. Sci.,
Sofia, Bulgaria

³ Department of Applied Analysis and Computational Mathematics,
Eötvös Loránd Univ., Budapest, Hungary

⁴ Department of Meteorology, Eötvös Loránd Univ., Budapest, Hungary

Abstract. Advection equations are an essential part of many mathematical models arising in different fields of science and engineering. It is important to treat such equations with efficient numerical schemes. The well-known Crank-Nicolson scheme will be applied. It will be shown that the accuracy of the calculated results can be improved when the Crank-Nicolson scheme is combined with the Richardson Extrapolation.

Keywords: Advection equations, Numerical methods, Crank-Nicolson scheme, Richardson Extrapolation.

1 One-Dimensional Advection Equations

Consider the advection equation:

$$\frac{\partial c}{\partial t} = -u \frac{\partial c}{\partial x}, \quad x \in [a_1, b_1] \subset (-\infty, \infty), \quad t \in [a, b] \subset (-\infty, \infty). \quad (1)$$

The wind velocity $u = u(x, t)$ is some given function. Equation (1) must always be considered together with appropriate initial and boundary conditions. The well-known Crank-Nicolson scheme (see, for example, [2, p. 63]) can be applied in the numerical treatment of (1). The computations are carried out by the following formula:

$$\begin{aligned} &\sigma_{i,n+0.5} c_{i+1,n+1} + c_{i,n+1} - \sigma_{i,n+0.5} c_{i-1,n+1} + \\ &+ \sigma_{i,n+0.5} c_{i+1,n} - c_{i,n} - \sigma_{i,n+0.5} c_{i-1,n} = 0 \end{aligned} \quad (2)$$

when the Crank-Nicolson scheme is used. The quantity $\sigma_{i,n+0.5}$ is defined by

$$\sigma_{i,n} = \frac{k}{4h} u(x_i, t_{n+0.5}) \quad (3)$$

where $t_{n+0.5} = t_n + 0.5k$ and the increments h and k of the spatial and time variables are introduced by using two equidistant grids:

$$G_x = \left\{ x_i, i = 0, \dots, N_x \mid x_0 = a_1, x_i = x_{i-1} + h, i = 1, \dots, N_x, h = \frac{b_1 - a_1}{N_x} \right\} \quad (4)$$

$$G_t = \left\{ t_n, n = 0, \dots, N_t \mid t_0 = a, t_n = t_{n-1} + k, n = 1, \dots, N_t, k = \frac{b-a}{N_t} \right\} \quad (5)$$

2 Application of the Richardson Extrapolation

Assume that a one-dimensional hyperbolic equation similar to (1) is treated by an arbitrary numerical method, which is of order $p \geq 1$ with regard to the two independent variables x and t . Let $\{z_{i,n+1}\}_{i=0}^{N_x}$ be the set of approximations of the solution of (1) calculated for $t = t_{n+1} \in G_t$ at all grid-points x_i , $i = 0, 1, \dots, N_x$, of G_x (4) by using the numerical method chosen and the corresponding approximations $\{z_{i,n}\}_{i=0}^{N_x}$ calculated at the previous time-step, i.e. for $t = t_n \in G_t$. Introduce vectors $\bar{c}(t_{n+1})$, $\bar{z}_n$ and $\bar{z}_{n+1}$ the components of which are $\{c(x_i, t_{n+1})\}_{i=0}^{N_x}$, $\{z_{i,n}\}_{i=0}^{N_x}$ and $\{z_{i,n+1}\}_{i=0}^{N_x}$ respectively. Since the order of the numerical method is assumed to be p with regard both to x and to t , we can write:

$$\bar{c}(t_{n+1}) = \bar{z}_{n+1} + h^p K_1 + k^p K_2 + O(k^{p+1}) \quad (6)$$

where K_1 and K_2 are some quantities, which do not depend on h and k . It is convenient to rewrite the last equality in the following equivalent form:

$$\bar{c}(t_{n+1}) = \bar{z}_{n+1} + k^p K + O(k^{p+1}), \quad K \stackrel{\text{def}}{=} \left(\frac{h}{k}\right)^p K_1 + K_2 \quad (7)$$

If h and k are sufficiently small, then the sum $h^p K_1 + k^p K_2$ will be a good approximation of the error in the calculated values of the numerical solution $\bar{z}_{n+1}$. If K is bounded, $|K| < \infty$, then $k^p K$ will also be a good approximation of the error of $\bar{z}_{n+1}$. This means that if we succeed to eliminate the term $k^p K$ in (7), then we shall obtain approximations of order $p + 1$. The Richardson Extrapolation can be applied in an attempt to achieve such an improvement of the accuracy. In order to apply the Richardson Extrapolation when (1) is treated by the Crank-Nicolson scheme it is necessary to introduce an additional grid:

$$G_x^2 = \left\{ x_i, i = 0, 1, \dots, 2N_x \mid x_0 = a_1, \right. \\ \left. x_i = x_{i-1} + \frac{h}{2}, i = 1, \dots, 2N_x, h = \frac{b_1 - a_1}{N_x} \right\} \quad (8)$$

Assume that approximations $\{w_{i,n}\}_{i=0}^{2N_x}$ (calculated at the grid-points of G_x^2 for $t = t_n \in G_t$) are available and perform two small steps with a stepsize $k/2$ to compute $\{w_{i,n+1}\}_{i=0}^{2N_x}$. Use only the components with even indices i ,

$i = 0, 2, 4, \dots, 2N_x$ to form vector $\tilde{w}_{n+1}$. The following equality holds for this vector when the quantity K is defined as in (7):

$$\bar{c}(t_{n+1}) = \tilde{w}_{n+1} + \left(\frac{k}{2}\right)^p K + O(k^{p+1}) \quad (9)$$

It is possible to eliminate the quantity K from (7) and (9) by applying the following linear combination: multiply (9) by 2^p and subtract (7) from the result. Thus we obtain:

$$\bar{c}(t_{n+1}) = \bar{c}_{n+1} + O(k^{p+1}), \quad \bar{c}_{n+1} \stackrel{\text{def}}{=} \frac{2^p \tilde{w}_{n+1} - \bar{z}_{n+1}}{2^p - 1} \quad (10)$$

The approximation $\bar{c}_{n+1}$, being of order $p+1$, will be more accurate than both $\bar{z}_{n+1}$ and $\tilde{w}_{n+1}$ when h and k are sufficiently small. The device used to construct $\bar{c}_{n+1}$ is called Richardson Extrapolation (introduced in [1]). If the partial derivatives up to order $p+1$ exist and are continuous, then one should expect (10) to produce more accurate results than those obtained by the underlying numerical method.

Remark 1. The rest terms in the formulae given in this section will in general depend on both h and k . However, it is clear that h can be expressed as a function of k by using (7) and (6) and this justifies the use only of k in all the rest terms.

Remark 2. No specific assumptions were made in this section, neither about the particular partial differential equation, nor about the numerical method used to solve it. This was done in order to demonstrate that the idea on which the Richardson Extrapolation is based is very general. However, it must be emphasized that in the remaining part of this paper it will always be assumed that (i) equation (1) is solved under the assumptions made in Section 1 and (ii) the underlying numerical algorithm used to handle it numerically is the second-order Crank-Nicolson scheme.

One should expect the combination of the Richardson Extrapolation and the Crank-Nicolson scheme to be a third-order numerical method. However, the actual result is much better, because the following theorem holds:

Theorem 1. *If $c(x, t)$ from (1) is continuously differentiable up to order five in both x and t , then the numerical method based on the Richardson Extrapolation and the Crank-Nicolson scheme is of order four.*

The Richardson Extrapolation can be implemented in four different manners depending on the way in which the computations at the next time-step, step $n+2$, will be carried out.

1. **Active Richardson Extrapolation:** Use $\bar{c}_{n+1}$ as initial value to compute $\bar{z}_{n+2}$. Use the set of values $\{w_{i,n+1}\}_{i=0}^{2N_x}$ as initial values to compute $\{w_{i,n+2}\}_{i=0}^{2N_x}$ and $\tilde{w}_{n+2}$.

2. **Passive Richardson Extrapolation:** Use $\bar{z}_{n+1}$ as initial value to compute $\bar{z}_{n+2}$. Use the set of values $\{w_{i,n+1}\}_{i=0}^{2N_x}$ as initial values to compute $\{w_{i,n+2}\}_{i=0}^{2N_x}$ and $\tilde{w}_{n+2}$.
3. **Active Richardson Extrapolation with linear interpolation on the finer spatial grid (8):** Use $\bar{c}_{n+1}$ as initial values to compute $\bar{z}_{n+2}$. Set $w_{2i,n+1} = c_{i,n+1}$ for $i = 0, 1, \dots, N_x$. Use linear interpolation to obtain approximations of the values of $w_{i,n+1}$ for $i = 1, 3, \dots, 2N_x - 1$. Use the updated set of values $\{w_{i,n+1}\}_{i=0}^{2N_x}$ as initial values to compute $\{w_{i,n+2}\}_{i=0}^{2N_x}$ and $\tilde{w}_{n+2}$.
4. **Active Richardson Extrapolation with third-order interpolation on the finer spatial grid (8):** Use $\bar{c}_{n+1}$ as initial value to compute $\bar{z}_{n+2}$. Set $w_{2i,n+1} = c_{i,n+1}$ for $i = 0, 1, \dots, N_x$. Use third-order Lagrangian interpolation polynomials to obtain approximations of $w_{i,n+1}$ for $i = 3, 5, \dots, 2N_x - 3$ and second-order Lagrangian polynomials to obtain approximations of $w_{i,n+1}$ for $i = 1$ and $i = 2N_x - 1$ (i.e. to calculate $w_{1,n+1}$ and $w_{2N_x-1,n+1}$). Use the updated set of values $\{w_{i,n+1}\}_{i=0}^{2N_x}$ as initial values to compute $\{w_{i,n+2}\}_{i=0}^{2N_x}$ and $\tilde{w}_{n+2}$.

The improvements obtained by applying (10) are not used in the further computations when the Passive Richardson Extrapolation is selected. These improvements are partly used in the calculations related to the large step (only to compute $\bar{z}_{n+2}$) when the Active Richardson Extrapolation is used. An attempt to exploit the more accurate values also in the calculation of $\bar{w}_{n+2} = \{w_{i,n+1}\}_{i=0}^{2N_x}$ is made in the last two implementations.

Information about the actual application of the third-order Lagrangian interpolation is given below. Assume that $w_{2i,n+1} = c_{i,n+1}$ for $i = 0, 1, \dots, N_x$, i.e. the improved (by the Richardson Extrapolation) solution on the coarser grid (4) is projected at the grid-points with even indices $0, 2, \dots, 2N_x$ of the finer grid (8). The interpolation rule used to get better approximations at the grid-points of (8) which have odd indices can be described by the following formula:

$$w_{i,n+1} = -\frac{3}{48} w_{i-3,n+1} + \frac{9}{16} w_{i-1,n+1} + \frac{9}{16} w_{i+1,n+1} - \frac{3}{48} w_{i+3,n+1}, \quad (11)$$

$$i = 3, 5, \dots, 2N_x - 3$$

Formula (11) is obtained by using a third-order Lagrangian interpolation for the case where the grid-points are equidistant and when an approximation at the mid-point x_i of the interval $[x_{i-3}, x_{i+3}]$ is to be found. Only improved values are involved in the right-hand-side of (11).

Formula (11) cannot be used to improve the values at the points x_1 and x_{N_x-1} . It is necessary to use second-order interpolation at these two points:

$$w_{1,n+1} = \frac{3}{8} w_{0,n+1} + \frac{3}{4} w_{2,n+1} - \frac{1}{8} w_{4,n+1}, \quad (12)$$

$$w_{N_x-1,n+1} = \frac{3}{8} w_{N_x,n+1} + \frac{3}{4} w_{N_x-2,n+1} - \frac{1}{8} w_{N_x-4,n+1}$$

3 Introduction of Three Numerical Examples

An oscillatory example (EXAMPLE 1). Assume that the following relationships hold:

$$\begin{aligned} a &= a_1 = 0, & b &= b_1 = 2\pi, & u(x, t) &= 0.5, \\ f(x) &= [100 + 99 \sin(10x)] * 1.4679 * 10^{12}. \end{aligned} \quad (13)$$

The exact solution of the problem defined by (13) is $c(x, t) = f(x - ut)$. Function $f(x)$ can be seen in Fig. 1 a).

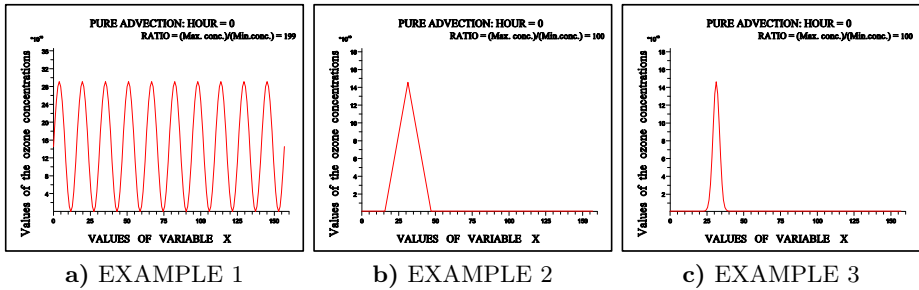


Fig. 1. The initial value conditions in the three examples. It is assumed that (i) there are 161 grid-points in the spatial interval and (ii) the initial values are ozone concentrations.

A discontinuous example (EXAMPLE 2). Another example is defined by the following relationships:

$$x \in [0, 50\,000\,000], \quad t \in [43\,200, 129\,600], \quad u(x, t) = 320 \text{ cm/s}. \quad (14)$$

The distance is measured in centimetres, which means that the length of the spatial interval is 500 kilometres. The time is measured in seconds (starting in the mid-night). This means that the calculations are started at 12 o'clock and finished at the same time in the next day. The initial values are given by

$$f(x) = 1.4679 * 10^{12} \quad \text{for } x \leq 5 * 10^6 \quad \text{or } x \geq 15 * 10^6, \quad (15)$$

$$f(x) = \left[1 + 99 * \frac{x - 5\,000\,000}{5\,000\,000} \right] * 1.4679 * 10^{12}, \quad 5 * 10^6 \leq x \leq 10 * 10^6, \quad (16)$$

$$f(x) = \left[1 + 99 * \frac{15\,000\,000 - x}{5\,000\,000} \right] * 1.4679 * 10^{12}, \quad 10 * 10^6 \leq x \leq 15 * 10^6. \quad (17)$$

The exact solution of the problem defined by (14) – (17) is given by $c(x, t) = f(x - u(t - 43200))$. The variation of function $f(x)$ defined by (15) – (17) can be seen in Fig. 1 b).

A smooth example with a sharp gradient (EXAMPLE 3). Assume that (14) holds and introduce:

$$f(x) = \left(1 + e^{-\omega(x-10\,000\,000)^2}\right) * 1.4679 * 10^{12}, \quad \omega = 10^{-12} \quad (18)$$

The exact solution of the problem defined by (14) and (18) is given by $c(x, t) = f(x - u(t - 43200))$. Function $f(x)$ from (18) can be seen in Fig. 1 c). Similar example was used in [4].

Similar advection module is a part of the large-scale air pollution model UNIDEM [3,5] and the quantities used in this section are either the same or very similar to the corresponding quantities in this model.

4 Numerical Results

In each experiment the first run is performed by using $N_t = 168$ and $N_x = 160$. Ten additional runs are performed after the first one. When a run is finished, both h and k are halved (this means that N_t and N_x are doubled) and a new run is started. Thus, in the eleventh run we have $N_t = 172032$ and $N_x = 163840$. Note too, that the ratio h/k is kept constant and, therefore K from (7) remains bounded as required in (8).

We are mainly interested in the behavior of the numerical error. The error is evaluated at the end of every hour (i.e. 24 times in each run) at the grid-points of the coarsest spatial grid in the following way. Assume that run number r , $r = 1, 2, \dots, 11$, is to be carried out and let $R = 2^{r-1}$. Then the error is calculated by

$$ERR_m = \max_{j=0,1,\dots,160} \left(\frac{|c_{\tilde{i},\tilde{n}} - c_{\tilde{i},\tilde{n}}^{exact}|}{\max(|c_{\tilde{i},\tilde{n}}^{exact}|, 1.0)} \right), \quad (19)$$

$$m = 0, 1, \dots, 24, \quad \tilde{i} = jR, \quad \tilde{n} = 7mR,$$

where $c_{\tilde{i},\tilde{n}}$ and $c_{\tilde{i},\tilde{n}}^{exact}$ are the calculated value and the reference solution at the end of hour m and at the grid-points of coarsest grid.

The global error made during the computations is estimated by using the following formula:

$$ERR = \max_{m=1,2,\dots,24} (ERR_m) \quad (20)$$

Numerical results obtained in the runs of the above three examples are given in Table 1 – 3.

Conclusions drawn by studying the results presented in Table 1:

- The Crank-Nicolson Scheme leads to second-order of accuracy when it is applied directly. This should be expected.
- The first three implementations of the Richardson Extrapolation (Active Richardson Extrapolation, Passive Richardson Extrapolation and Richardson

Table 1. Running the oscillatory advection example (EXAMPLE 1) by using the Crank-Nicolson Scheme directly and in combination with four versions of the Richardson Extrapolation. The convergence rate is given in brackets for the last method.

No	NT	NX	C-N only	Richardson Extrapolation [error (conv. rate)]			
				Active	Passive	Lin. interp.	3^{rd} order interp.
1	168	160	7.85E-01	2.04E-01	2.79E-01	3.83E-01	1.56E-02
2	336	320	2.16E-01	4.95E-02	7.14E-02	1.19E-01	1.23E-03 (12.7)
3	672	640	5.32E-02	1.25E-02	1.76E-02	2.47E-02	1.07E-04 (11.4)
4	1344	1280	1.33E-02	3.15E-03	4.33E-03	6.25E-03	1.15E-05 (9.3)
5	2688	2560	3.32E-03	7.87E-04	1.07E-03	1.57E-03	1.19E-06 (9.6)
6	5376	5120	8.30E-04	1.97E-04	2.67E-04	3.92E-04	1.48E-07 (8.1)
7	10752	10240	2.08E-04	4.92E-05	6.66E-05	9.81E-05	1.62E-08 (9.1)
8	21504	20480	5.19E-05	1.23E-05	1.66E-05	2.45E-05	1.96E-09 (8.2)
9	43008	40960	1.30E-05	3.08E-06	4.15E-06	6.13E-06	2.39E-10 (8.2)
10	86016	81920	3.24E-06	7.96E-07	1.04E-06	1.53E-06	3.24E-11 (7.4)
11	172032	163840	8.10E-07	1.92E-07	2.60E-07	3.83E-07	1.27E-11 (2.7)

Extrapolation with linear interpolation of the values of $\{w_{i,n+1}\}_{i=0}^{2N_x}$ on the grid-points of the finer spatial grid) lead also to second-order accuracy (instead of the fourth-order accuracy which should be expected). On the other hand, these three methods give more accurate results than those obtained by using directly the Crank-Nicolson Method.

- The combination of the Crank-Nicolson Scheme with the Richardson Extrapolation performs as a third-order numerical method when it is enhanced with third-order Lagrangian interpolation polynomials for improving the accuracy of the values of $\{w_{i,n+1}\}_{i=0}^{2N_x}$ on the finer spatial grid. Theorem 1 tells us that the combined method should be of order four. The lower accuracy achieved here is probably due to the use of interpolation of lower degree in formula (12).

Conclusions drawn by studying the results presented in Table 2:

- All five numerical methods (the direct implementation of the Crank-Nicolson Scheme and the four implementations of the Richardson Extrapolation) lead to first-order of accuracy. This probably should be expected (because of the presence of discontinuities).
- The four implementations of the Richardson Extrapolation give more accurate results than those obtained by using directly the Crank-Nicolson Scheme.
- The combination of the Crank-Nicolson Scheme with the Richardson Extrapolation performs best when it is enhanced with third-order Lagrangian interpolation polynomials for improving the accuracy of the values of $\{w_{i,n+1}\}_{i=0}^{2N_x}$ on the finer spatial grid. However, the improvements achieved are very modest also in this case.

Table 2. Running the example with discontinuous derivatives (EXAMPLE 2) by using the Crank-Nicolson Scheme directly and in combination with four versions of the Richardson Extrapolation. The convergence rate is given in brackets for the last method.

No	NT	NX	C-N only	Richardson Extrapolation [error (conv. rate)]			
				Active	Passive	Lin. interp.	3^{rd} order interp.
1	168	160	1.34E-01	7.67E-02	7.93E-02	1.17E-01	4.98E-02
2	336	320	7.69E-02	4.42E-02	4.57E-02	6.66E-02	2.76E-02 (1.80)
3	672	640	4.42E-02	2.55E-02	2.56E-02	3.99E-02	1.55E-02 (1.78)
4	1344	1280	2.55E-02	1.64E-02	1.57E-02	2.45E-02	8.57E-03 (1.81)
5	2688	2560	1.64E-02	1.06E-02	1.07E-02	1.51E-02	4.59E-03 (1.87)
6	5376	5120	1.06E-02	5.80E-03	5.89E-03	9.68E-03	2.32E-03 (1.98)
7	10752	10240	5.80E-03	3.40E-03	4.09E-03	5.51E-03	1.19E-03 (1.95)
8	21504	20480	3.40E-03	2.35E-03	2.48E-03	3.23E-03	6.58E-04 (1.81)
9	43008	40960	2.35E-03	1.33E-03	1.10E-03	2.26E-03	2.38E-04 (2.75)
10	86016	81920	1.33E-03	9.29E-04	9.45E-04	1.14E-03	1.50E-04 (1.59)
11	172032	163840	9.36E-04	4.08E-04	2.99E-04	8.88E-04	2.79E-05 (4.94)

Table 3. Running the smooth advection example (EXAMPLE 3) by using the Crank-Nicolson Scheme directly and in combination with four versions of the Richardson Extrapolation. The convergence rate is given in brackets for the last method.

No	NT	NX	C-N only	Richardson Extrapolation [error (conv. rate)]			
				Active	Passive	Lin. interp.	3^{rd} order interp.
1	168	160	7.37E-01	3.99E-01	3.78E-01	6.41E-01	1.45E-01
2	336	320	4.00E-01	1.27E-01	1.00E-01	3.34E-01	1.74E-02 (8.4)
3	672	640	1.25E-01	3.08E-02	1.28E-02	1.09E-01	1.22E-03 (14.2)
4	1344	1280	3.08E-02	7.76E-03	9.07E-04	2.67E-02	1.73E-05 (15.8)
5	2688	2560	7.77E-03	1.95E-03	5.37E-05	6.84E-03	4.84E-06 (16.0)
6	5376	5120	1.95E-03	4.89E-04	3.30E-06	1.72E-03	3.03E-07 (16.0)
7	10752	10240	4.89E-04	1.22E-04	2.07E-07	4.30E-04	1.89E-08 (16.0)
8	21504	20480	1.22E-04	1.23E-05	1.29E-08	1.07E-04	1.18E-09 (16.0)
9	43008	40960	3.09E-05	7.65E-06	8.09E-10	2.69E-05	7.61E-11 (15.5)
10	86016	81920	7.65E-06	1.91E-07	5.06E-11	6.72E-06	9.85E-12 (7.7)
11	172032	163840	1.91E-06	4.78E-07		1.68E-07	4.97E-12 (2.0)

Conclusions drawn by studying the results presented in Table 3:

- The direct application of the Crank-Nicolson Scheme leads to quadratic convergence.
- The active Richardson Extrapolation and the Richardson Extrapolation based on the use of linear interpolation behave as second order methods, but give slightly better accuracy than that obtained when the Crank-Nicolson scheme is applied directly.
- The Passive Richardson Extrapolation behaves as method of order four for this example.
- The fourth implementation of the Richardson Extrapolation behaves as a numerical method of order four. This result is in agreement with the statement of Theorem 1. We should mention that the interpolation formula (12) for the spatial boundary grid-points gives very accurate approximations for this particular example.

Acknowledgements

This research is supported in parts by the Bulgarian NSF grants DTK 02-44/2009, DO 02-115/2008 and DO 02-161/2008. The Centre for Supercomputing at the Technical University of Denmark gave us access to several powerful parallel supercomputers for running the experiments related to this study.

References

1. Richardson, L.F.: The deferred approach to limit, I - Single lattice. Philosophical Transactions of the Royal Society, London, Ser. A 226, 299–349 (1927)
2. Strikwerda, J.C.: Finite Difference Schemes and Partial Differential Equations. SIAM, Philadelphia (2004)
3. Zlatev, Z.: Computer Treatment of Large Air Pollution Models. Kluwer Academic Publishers, Dordrecht (1995)
4. Zlatev, Z., Berkowicz, R., Prahm, L.P.: Testing Subroutines Solving Advection-Diffusion Equations in Atmospheric Environments. Computers and Fluids 11, 13–38 (1983)
5. Zlatev, Z., Dimov, I.: Computational and Numerical Challenges in Environmental Modelling. Elsevier, Amsterdam (2006)

Programming Problems with a Large Number of Objective Functions

Cornel Resteanu¹ and Romica Trandafir²

¹ National Institute for Research and Development in Informatics
8-10 Averescu Avenue, 011455, Bucharest 1, Romania
`resteanu@ici.ro`

² Technical University of Civil Engineering
24 Lacul Tei Avenue, 020396, Bucharest, Romania
`romica@utcb.ro`

Abstract. The paper treats the Multi-Objective Programming problem with a large composite set of (linear and nonlinear) objective functions, the domain of feasible solutions being defined by a set of linear equalities/inequalities representing a large scale problem. One constructs a preferred solution i.e. a non-dominated solution chosen via extending the decision-making framework. A feasible approach, for this class of problems, is to use a solver for the Linear Programming problems and a solver for Multiple Attribute Decision Making problems in combination with Parallel and Distributed Computing techniques based on a GRID configuration.

Keywords: Multi-Objective Programming, Large Scale Problems, Multiple Attribute Decision Making, Parallel and Distributed Computing, GRID Nets.

1 Introduction

For the Multi-Objective Programming (MOP) problems [1], when the objective functions are linear and the domain of feasible solutions is defined by a set of linear equalities/inequalities, theoretically there are two essential different solving methods based on SIMPLEX or interior point algorithms.

The first method iterates step by step mono-objective problems upon each objective function but at each step enriching the domain of the feasible solutions with a new inequality constructed by the previous objective function forced to be, conform with the nature of optimum, greater or smaller than the optimal value obtained in the last optimization, often relaxed with a convenient small value. The method takes a lot of time because the re-optimization after adding a new line to the set of linear equalities/inequalities which define the domain of the feasible solutions is time consuming. Moreover, in most of cases, the repetitive process must be re-taken from an incipient step because the objective functions' values are inadequate for user's goal.

The second method also iterates step by step mono-objective problems upon each objective function but always maintaining the initial domain of the feasible

solutions. Considering the optimal values of the objective functions as an ideal goal to reach, a final minimization of an objective function expressing the distance to this ideal is made. The method takes less time than the first one but in most cases the final solution is unacceptable from the practical point of view, even in the construction of final objective function one works with different distances and different filtration of non-dominate solutions techniques.

If nonlinear functions (beside linear functions) appear among the objectives, the above described methods do not work. If the objective functions are in a large number and over a solutions domain representing a large scale problem, the solving becomes difficult and can be practically approached only using modern techniques as in the following.

2 MOP Problem Presentation

The class of MOP problems broached in this paper, presented in a synthetic, vector form, is:

$$\begin{cases} \max \underline{w}^r \underline{f}(\underline{x}) \\ \underline{g}(\underline{x}) \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} \underline{b}, \\ \underline{x} \geq \underline{0} \end{cases} \quad (1)$$

with the following specifications:

$$\underline{x} = (x_1, x_2, \dots, x_n), \quad x_i \geq 0, \quad (\forall) i \in \overline{1, n} \quad (n \text{ decision variables});$$

$$\underline{f}(\underline{x}) = (f_1(\underline{x}), \dots, f_k(\underline{x}), f_{k+1}(\underline{x}), \dots, f_l(\underline{x})) \quad (l \text{ objective functions}),$$

where:

$$f_o(\underline{x}) = \sum_{i=1}^n c_{oi} x_i, \quad c_{oi} \in \mathfrak{R}, \quad (\forall) o \in \overline{1, k}, \quad (\forall) i \in \overline{1, n} \quad (k \text{ linear functions})$$

$$f_o(\underline{x}) = n \ln o(\underline{x}), \quad (\forall) o \in \overline{k+1, l} \quad (l - k \text{ increasing nonlinear functions});$$

$$\underline{w} = (w_1, \dots, w_k, w_{k+1}, \dots, w_l), \quad 1 \geq w_1 \geq \dots \geq w_k \geq$$

$$\geq w_{k+1} \geq \dots \geq w_l \geq 0, \quad \sum_{o=1}^l w_o = 1$$

(l objective functions' weights, decreasing with the accorded importance);

$$\underline{g}(\underline{x}) = (g_1(\underline{x}), \dots, g_m(\underline{x})) \quad (m \text{ constraints})$$

where

$$g_j(\underline{x}) = \sum_{i=1}^n a_{ji} x_i \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_j, \quad j \in \overline{1, m},$$

$$a_{ji}, b_j \in \mathfrak{R}, (\forall) i \in \overline{1, n}, j \in \overline{1, m}.$$

Thus the problem (1) becomes problem (2):

$$\left\{ \begin{array}{l} \max[w_1 \sum_{i=1}^n c_{1i} x_i, \dots, w_k \sum_{i=1}^n c_{ki} x_i, w_{k+1} nlin_{k+1}(x_1, \dots, x_n), \dots, \\ \quad \dots, w_l nlin_l(x_1, \dots, x_n)] \\ \sum_{i=1}^n a_{ji} x_i \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_j, j = 1, \dots, m \\ x_i \geq 0, i = 1, \dots, n. \end{array} \right. \quad (2)$$

One considers the first k linear functions as main objectives and the rest $l - k$ increasing nonlinear functions as secondary objectives. As numbers, $k \ll l - k$, k is of tens order and $l - k$ is of hundreds order or even of thousands order.

Definitions and specifications

1. Let be (2') the problem (2) restrained to linear objective functions.
2. $\underline{x}, \underline{y} \in \mathfrak{R}^n : \underline{x} \leq \underline{y} \Leftrightarrow x_i \leq y_i, (\forall) 1 \leq i \leq n$.
3. $X = \left\{ \underline{x} | \underline{g}(\underline{x}) \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} \underline{b} \right\}$ is the set of *feasible solutions* of problem (2').
4. $S = \{\underline{f}(\underline{x}) | \underline{x} \in X\}$ is, in problem (2'), the set of *$\underline{f}(\underline{x})$ -vector function's values*.
5. $\underline{x}^*$ is the *optimal solution* for problem (2') $\Leftrightarrow \underline{x}^* \in X$ and $f_o(\underline{x}^*) \geq f_o(\underline{x})$, $(\forall) \underline{x} \in X$ and $(\forall) o \in \overline{1, k}$.
6. $\underline{x}^*$ is a *non-dominated solution* for problem (2') $\Leftrightarrow$ there are not $\underline{x} \in X$ so that $f_o(\underline{x}^*) \leq f_o(\underline{x})$, $(\forall) o \in \overline{1, k}$ and $f_o(\underline{x}^*) < f_o(\underline{x})$ for at least one o .
7. For the MOP problem (2), the *restrained decision-making framework* refers to the k linear objective functions, their values and of course the solutions.
8. For the MOP problem (2), the *extended decision-making framework* refers to the all l objective functions, their values and of course the solutions.
9. $\underline{x}^*$ is a *preferred solution* for problem (2) $\Leftrightarrow$ is non-dominated and it is chosen by decision maker via extending decision-making framework.

3 Gridification of (2) Class MOP Problem

A tool to treat this problem with the composite and large set of objective functions would be necessary, not in the case of small size problems, but for large scale problems. It is well-known that such a tool does not exist on the informatics market. Therefore one should envisages a method for this purpose, using the

existent tools. A feasible approach, in this particular case, is a software platform developed by a solver for the Linear Programming (LP) problems [2] and a solver for Multiple Attribute Decision Making (MADM) problems [3] in combination with Parallel and Distributed Computing (PDC) techniques [4, 5, 6, 7] based on a GRID [8] configuration. The hardware platform, a multitude of computers, belonging to a GRID structure, has the same importance as the software platform because only by having performance in both directions it is possible to have a global performance.

Europe had a leading grid computing project called Enabling Grids for E-science (EGEE 2002-2004 and EGEE 2004-2010) which provided a computing support infrastructure for over 10,000 researchers world-wide and opened the possibility for the European Grid Infrastructure organization to coordinate the National Grid Initiatives. To support the development of the Romanian Grid projects, as a consistent and coherent part of the European R&D activity in this field, was set up the Rogrid consortium. This consortium has as partners: National Institute for Research and Development in Informatics (ICI Bucharest.), University Politehnica of Bucharest (UPB), National Institute for Physics and Nuclear Engineering (IFIN-HH), National Institute for Aerospace Research (INCAS), University of Bucharest (UB), Technical University of Cluj-Napoca (UTCN), Western University of Timisoara (WUT). It is responsible for the implementation of the National Grid Infrastructure and for the development of the software that will run on the above mentioned infrastructure. Therefore, the algorithm described below and its grid implementation were analyzed and then approved by Rogrid specialists.

First, one defines the MOP solving configuration as a sub-net in the national grid net. The sub-net has in its nodes a powerful server and a number of multi-processors. This number is equal to the linear objective functions' number. If the server remains the same after the pervasive solving service [9] has been installed on sub-net, the rest of multi-processors may differ from a solving instance to another solving instance. On server are assigned the mathematical modeling operations, i.e. stocking the mathematical model in a format that facilitates the operations of adding / modifying / deleting the model entities, the generating of MOP problem, the running of the first phase of the basic solving algorithm and the construction of the preferred solution. On the rest of multi-processors are assigned the optimizations upon the linear objective functions, in re-optimization regime, which are parallel operations scheduled by server. In fact, this configuration creates the possibility for two kinds of parallelism. Obviously, the server can be accessed at the same time by a large number of users, the Apache and MySQL corresponding parameters are fixed to 1000, greater than the real necessity. Such type of parallelism is called users concurrency and is provided by the basic software without any effort wasted by the designers or programmers. On the contrary, the parallelism and the distribution for solving a MOP problem is the designers and programmers achievement.

When a user wants to solve such kind of problem, the following algorithm is performed:

Step 0. On server:

Considering that the problem is already in the data base, transfer, from the data base, the MPS (Mathematical Programming System) form of (2') on the server's disk as a current solving problem. Register, in the current solving problems' file, the identity data of the current solving problem and the server start time. Inspect the GRID to find k multi-processors $(mp_1, mp_2, \dots, mp_k)$, slightly or at all charged with other works. Do a multi-processors' permutation such that the less charged will be the first one $(mp_{\sigma(1)}, mp_{\sigma(2)}, \dots, mp_{\sigma(k)})$.

Step 1. On server:

Through the agency of the LP problem solver, transform the MPS form into so call work standard form:

- define the *slack* and *excess* variables, namely d_j^- and d_j^+ ,

$$d_j^- = \begin{cases} d_j^- = b_j - g_j(\underline{x}) & \text{if } b_j \geq g_j(\underline{x}) \\ 0 & \text{otherwise} \end{cases}$$

$$d_j^+ = \begin{cases} d_j^+ = g_j(\underline{x}) - b_j & \text{if } b_j \leq g_j(\underline{x}) \\ 0 & \text{otherwise} \end{cases}$$

- construct, from (2'), the following problem:

$$\begin{cases} \min \sum_{j=1}^m [d_j^- + d_j^+] \\ g_j(\underline{x}) + d_j^- - d_j^+ = b_j, \quad j=1, \dots, m \\ x_j \geq 0, \quad i=1, \dots, n, \quad \underline{d}^-, \underline{d}^+ \geq \underline{0}, \quad d_j^- \cdot d_j^+ = 0, \quad (\forall) j=1, \dots, m. \end{cases} \quad (3)$$

Do solve the problem (3). The solving of problem (3) shows whether X , the set of feasible solutions for problem (2), is not void. That happens if the objective function of (3) is null. If X is void, do transfer the slack and excess values to server for analysis and model updating and stop the procedure removing the current problem from the current solving problems' file else go to next step.

Step 2. On $mp_{\sigma(1)}, mp_{\sigma(2)}, \dots, mp_{\sigma(k)}$, in parallel:

Transfer, from server to $mp_{\sigma(p)}$, $p \in \overline{1, k}$,

$$\begin{cases} \max \left[\sum_{i=1}^n c_{pi} x_i \right] \\ \sum_{i=1}^n a_{ji} x_i \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_j, \quad j=1, \dots, m \\ x_i \geq 0, \quad i=1, \dots, n, \end{cases} \quad (4)$$

the problem (4) in the MPS standard form, having as objective function $f_p(\underline{x}) = \sum_{i=1}^n c_{pi}x_i$ and with the optimal base obtained at (3)-problem's solving, which is an starting optimal base for (4).

Step 3. On $mp_{\sigma(1)}, mp_{\sigma(2)}, \dots, mp_{\sigma(k)}$, in parallel:

Solve, in the re-optimization regime, k (4)-problems, for each $p \in \overline{1, k}$, resulting $\underline{x}^{*p}$, $f_p^* = f_p(\underline{x}^{*p})$. Let be $\underline{f}^* = (f_1^*, f_2^*, \dots, f_k^*)$ the ideal solution.

When a couple $\underline{x}^{*p}$, $f_p^* = f_p(\underline{x}^{*p})$, $p \in \overline{1, k}$, is ready, transfer it to server. Wait from the last $p \in \overline{1, k}$ solving and transfer process. If $(\exists)\underline{x}^* \in X$ with $\underline{f}(\underline{x}^*) = \underline{f}^* \Rightarrow \underline{x}^*$ is the optimal solution and the algorithm stops because an optimal solution is better than a preferred solution. However, usually $(\exists)\underline{x}^* \notin X$. In this case, a preferred solution will be constructed via extending the decision-making framework, namely taking into consideration the information kept in the rest of objective functions.

Step 4. On server:

Using $y_{po} = f_o(\underline{x}^{*p})$, $(\forall)p \in \overline{1, k}$, $(\forall)o \in \overline{1, l}$ one constructs the extended consequence matrix as in the following table:

Table 1. Extended consequences matrix

	$f_1 \dots f_k$	$f_{k+1} \dots f_l$
$\underline{x}^{*1}$	$y_{11} \dots y_{1k}$	$y_{1k+1} \dots y_{1l}$
$\underline{x}^{*2}$	$y_{21} \dots y_{2k}$	$y_{2k+1} \dots y_{2l}$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\underline{x}^{*k}$	$y_{k1} \dots y_{kk}$	$y_{kk+1} \dots y_{kl}$

Let be $m_o^- = \min_{p \in \overline{1, k}} f_o(\underline{x}^p)$, $m_o^+ = \max_{p \in \overline{1, k}} f_o(\underline{x}^p)$, $(\forall)o \in \overline{1, l}$ and therefore one obtains: $m^- = (m_1^-, m_2^-, \dots, m_l^-)$, the negative ideal point and $m^+ = (m_1^+, m_2^+, \dots, m_l^+)$, the positive ideal point. These values, together with the weights $\underline{w} = (w_1, \dots, w_k, w_{k+1}, \dots, w_l)$ are the entry data for a MADM method [10, 11, 12, 13], in this case the TOPSIS method, that offers a weighting for $\underline{x}^{*1}, \underline{x}^{*2}, \dots, \underline{x}^{*k}$ and implicitly for $f_1, f_2, \dots, f_k$ through the agency of a weights vector $\underline{w}^* = (w_1^*, w_2^*, \dots, w_k^*)$ as in the following:

The distances to the negative and positive ideal points are:

$$d^-[\underline{x}^{*p}] = \left(\sum_{o=1}^l (m_o^- - y_{po})^2 \right)^{1/2}, \quad d^+[\underline{x}^{*p}] = \left(\sum_{o=1}^l (m_o^+ - y_{po})^2 \right)^{1/2} \quad (\forall)p \in \overline{1, k}.$$

One calculates $d_p^* = \frac{d^-[\underline{x}^{*p}]}{d^-[\underline{x}^{*p}] + d^+[\underline{x}^{*p}]}$, $(\forall)p \in \overline{1, k}$ and consider:

$$\frac{d_1}{w_1^*} = \frac{d_2}{w_2^*} = \dots = \frac{d_k}{w_k^*} \frac{\sum_{p=1}^k d_p}{\sum_{p=1}^k w_p^*} \text{ with } \sum_{p=1}^k w_p^* = 1.$$

From this one obtains the weights $w_o^* = d_o / \sum_{p=1}^k d_p$, $o \in \overline{1, k}$. The weight w_o^* of one linear objective function f_o , $(\forall)o \in \overline{1, k}$ one interprets as power to attract good values for the nonlinear objective functions.

Step 5. From (3) one constructs and solves, with the last basis, i.e. in the re-optimization regime, the problem:

$$\left\{ \begin{array}{l} \max \left[w_1^* \sum_{i=1}^n c_{1i} x_i + \dots + w_k^* \sum_{i=1}^n c_{ki} x_i \right] \\ \sum_{i=1}^n a_{ji} x_i \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_j, \quad j = 1, \dots, m \\ x_i \geq 0, \quad i = 1, \dots, n \end{array} \right. \quad (5)$$

resulting $\underline{x}^{**}$, $f^{**}(\underline{x}^{**})$, the final solution.

Step 6. The final solution is transferred in the data base, the total solving time is computed and the current problem is removed from the current solving problems' file.

It is to notice that the algorithm is not interactive, therefore the solution, in the case of non void solutions domain, is delivered automatically.

4 Conclusions

Web Enabled Optimization [14] is a new trend in treating the *Operation Research* problems. Moreover, the GRID facilities' use in approaching the concurrent, parallel and distributed optimizations is the last decade novelty in the field. The content of this paper may be considered a terse and telling example for the above assertions.

Optimizations like in this paper encourage the *Advanced E-Applications* fast development. In the authors' opinion, elaborate *Web-Optimization Applications* [15, 16, 17, 18] will soon become facts on the IT market.

References

1. Hwang, C.L., Masud, A.S.M.: Multiple Objective Decision Making. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin (1979)
2. <http://www.faqs.org/faqs/linear-programming-faq/>,
<ftp://softlib.cs.rice.edu/pub/>
3. Resteanu, C., Șomodi, M., Andreica, M., Mitan, E.: Distributed and parallel computing in MADM domain using the OPTCHOICE software. In: Proceedings of the 7th WSEAS International Conference on Applied Computer Science (ACS 2007), Venice, Italy, November 21-23, pp. 376–384 (2007)
4. Grama, A., Karpis, G., Kumar, V., Gupta, A.: Introduction to Parallel Computing: Design and Analysis of Parallel Algorithms. Addison Wesley, Reading (2003)
5. Jordan, H.F., Alagband, G., Jordan, H.F.: Fundamentals of Parallel Processing. Prentice Hall, Englewood Cliffs (2002)
6. Dongarra, J., Madsen, K., Wasniewski, J. (eds.): Applied Parallel Computing: State of the Art in Scientific Computing, 1st edn., April 11. LNCS. Springer, Heidelberg (2006)
7. Quinn, M.: Parallel Programming in C with MPI and OpenMP, 1st edn. McGraw-Hill Science / Engineering / Math (2003)
8. Silva, V.: Grid Computing for Developers. Programming Series, p. 547. Charles River Media, Hingham (2005) ISBN 1584504242
9. <http://www.wordreference.com/definition/pervasive>
10. Hwang, C.L., Yoon, K.: Multiple Attribute Decision Making. Springer, Berlin (1981)
11. Hwang, C.L., Lin, M.J.: Group Decision Making under Multiple Criteria. Springer, Heidelberg (1997)
12. Resteanu, C., Filip, F.G., Ionescu, C., Șomodi, M.: On Optimal Choice Problem Solving. In: Sage, A.P., Zheng, W. (eds.) Proceedings of SMC 1996 Congress, Beijing, October 14-17, pp. 1864–1869. IEEE Publishing House, Piscataway (1996)
13. Resteanu, C.: MADM Theory and practice. In: Ed. ICI (2006) (in Romanian)
14. <http://www-neos.mcs.anl.gov/>
15. Cohen, M.-D., Kelly, C.B., Medaglia, A.L.: Decision Support Systems with Web-Enabled software. Interfaces 31(2), 109–129 (2001)
16. Teodorescu, H.-N., Zbancioc, M.-D., Pistol, L.: Parallelizing Neuro-fuzzy Economic Models in a GRID Environment. Studies in Informatics and Control 17(1), 5–16 (2008)
17. Petcu, D., Macariu, G., Carstea, A., Frincu, M.E.: Service-Oriented Symbolic Computing. In: Antonopoulos, N., Exarchakos, G., Liotta, A., Li, M. (eds.) Handbook of Research on P2P and Grid Systems for Service-Oriented Computing - Models, Methodologies and Applications, Information Science Reference, ch. 15, pp. 1053–1075 (January 2010) ISBN: 978-1-61520-686-5
18. Petcu, D., Iordan, V.: Understanding Service Oriented Architectures in the Classroom - from Web Services to Grid Services. In: Papadopoulos, G.A., Wojtkowski, W., Wojtkowski, G., Wrycza, S., Zupancic, J. (eds.) Procs. ISD 2008 - Information Systems Development Towards a Service Provision Society, pp. 831–838. Springer Hardcover, Heidelberg (2009) ISBN: 978-0-387-84809-9

First Results of SEE-GRID-SCI Application CCIAQ

Dimitar Syrakov¹, Valery Spiridonov¹, Kostadin Ganey², Maria Prodanova¹,
Andrey Bogachev¹, Nikolai Miloshev², and Kiril Slavov¹

¹ National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences,
Sofia, Bulgaria

`iglika.etropolska@meteo.bg`

² Geophysical Institute, Bulgarian Academy of Sciences, Sofia, Bulgaria
`kganey@geophys.bas.bg`

Abstract. Intensive long-term meteorological modeling took place over an area covering Bulgaria with resolution of 10 km. The climatic version of the operational weather forecast model ALADIN was applied for simulating 3 time slices: 1960-2000, 2020-2050 and 2070-2100, following the IPCC scenario A1B. The differences of climatic fields for the 3 periods are presented and interpreted. The created met-data base is used to estimate the impact of climate changes on air quality, as well. A respective modeling System was created on the base of US EPA Models-3 tool (MM5, CMAQ and SMOKE). Calculations for the last 10 years of each time slice are performed. Grid technology in the frame of SEE-GRID-SCI project is used to perform this enormous volume of calculations as an application abbreviated to CCIAQ (Climate Change Impact on Air Quality). The results are presented and interpreted in the study.

1 Introduction

Many scientific projects and publications are aimed to assessment of the possible climate changes and on its impact on various areas of human activity and environment. It is not possible in such a short paper to give, at least, a brief overview of the current state-of-the-art of the problem. The EC FP6 project CECILIA (<http://www.cecilia-eu.org/>) is only one of the related projects. CECILIAs Work Packages 1 and 2 are connected with long-term meteorological simulations aiming creation of respective databases of high resolution (10 km) meteorological fields capable to retrieve climate estimates. WP7 aims at long-term simulations of chemistry air quality models driven by Regional Climate Models (RCMs) for present climate and for future projections with fine resolution of 10 km for some target regions of Central and Eastern Europe. The chemical boundary conditions for these regions are prepared by calculations covering the whole of Europe with coarser grid resolution of 5050 km.

2 Meteorological Modeling and Meteorological Data Base, Verification

The meteorological data base for this study is created off-line by the ALADIN Regional Climatic Model. It is a modification of the current operative weather forecast system in the National Institute of Meteorology and Hydrology (NIMH) of Bulgaria. Its creation is a result of a project between France and Bulgaria [7, 8, 9]. ALADIN-Climate is a regional version of ARPEGEClimate [2]. ARPEGE and ALADIN use the same executables (same dynamical core, same physical parameterizations).

The simulations span two scenario time slices 2021-2050 (Near Future, NF) and 2071-2100 (Far Future, FF). They are driven at the lateral boundaries by meteorological fields from a corresponding global simulation with the ARPEGE model under forcing from the SRES-A1B IPCC greenhouse gas IPCC scenario [4]. The validation period is 1961-1990 (Control Run, CR) with ERA40 [14] forcing as boundary conditions.

The ALADIN output is a binary file with 6-hour time resolution transformed to standard GRIB-format. The created by ALADIN-Climate meteorological data base consists of such 6-hour GRIB files containing the main meteorological parameters in 31 standard p-levels going up to 50 hPa.

The last time slice calculation results are used for validation by comparing the respective model climate with present climate estimates (1960-2000). Daily precipitation and temperature are taken from 56 stations and respectively averaged in space and time. The quality of the simulation is presented in Fig. 1.

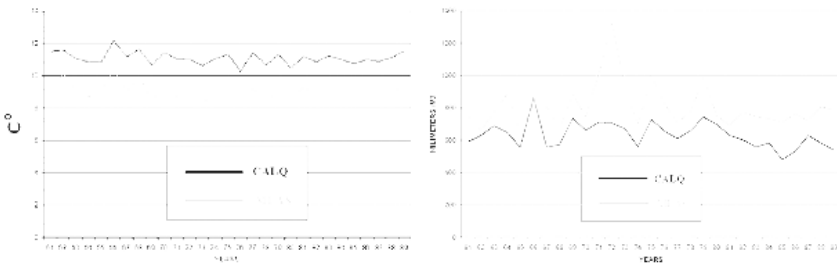


Fig. 1. Comparison between calculated and measured yearly temperatures (left) and precipitation (right) over 56 stations in Bulgaria for 1961-1990

3 Climate Change in the Temperature and the Precipitation

The verification shows that the errors are relatively permanent in the period, i.e. they can be considered as systematical. The difference between the simulated future period and the present one gives the tendencies in the future,

i.e. the climate change dimensions for the domain used. This difference eliminates in big extent the systematical error of ALADIN. The two future periods (NF and FF) are simulated; tendencies presented in Fig. 2, 3 for temperature and precipitation, respectively.

The temperature shows positive tendency of increasing of 2°C in NF and 3.7°C in FF. There could not be mentioned significant differences of the temperature distribution in the integration domain. The precipitations tendencies show steady signal for decreasing of 15-20 % comparing with the reference run in the South-East part of the domain and North Greece. The decreasing of precipitation is approximately the same in both simulations for NF and FF.

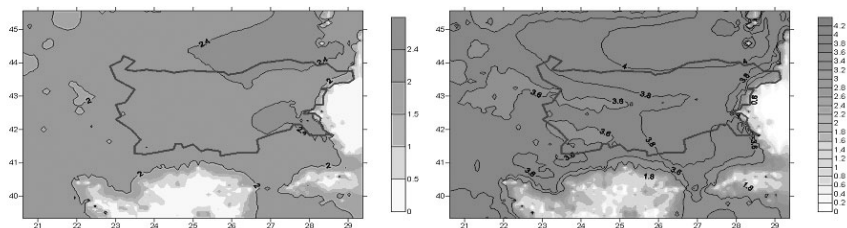


Fig. 2. Tendency in the mean temperature in $^{\circ}\text{C}$: NF-CR (left) and FF-CR (right)

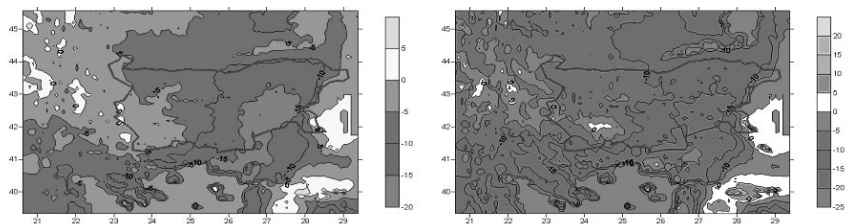


Fig. 3. Tendency in the yearly accumulated precipitation in %: $100 \cdot (\text{NF-CR}) / \text{CR}$ (left) and $100 \cdot (\text{FF-CR}) / \text{CR}$ (right)

4 Air Pollution Modeling and Air Quality Database, Verification

For assessing the climate change impact on air quality, a modeling System was elaborated based on the open source US EPA Models-3 air quality modeling tool, consisting of:

- CMAQ (<http://www.cmaq-model.org/>), Community Multi-scale Air Quality model, being the chemical-transport model of the System;

- MM5 (<http://box.mmm.ucar.edu/mm5/>), The 5th generation PSU/NCAR Meso-Meteorological Model, used as meteorological pre-processor to CMAQ; and
- SMOKE (<http://www.smoke-model.org/>), Sparse Matrix Operator Kernel Emissions modeling system, being the emission pre-processor to CMAQ.

A number of interfaces (Linux scripts and FORTRAN codes) are created as to link those models with different types input information in a chain capable to perform long term calculations. The calculations are performed for a region containing mainly Bulgaria with 10×10 km resolution nested in the ALADIN domain. Here, a short description of the System will be presented. One can find more detailed description in [11, 12, 13].

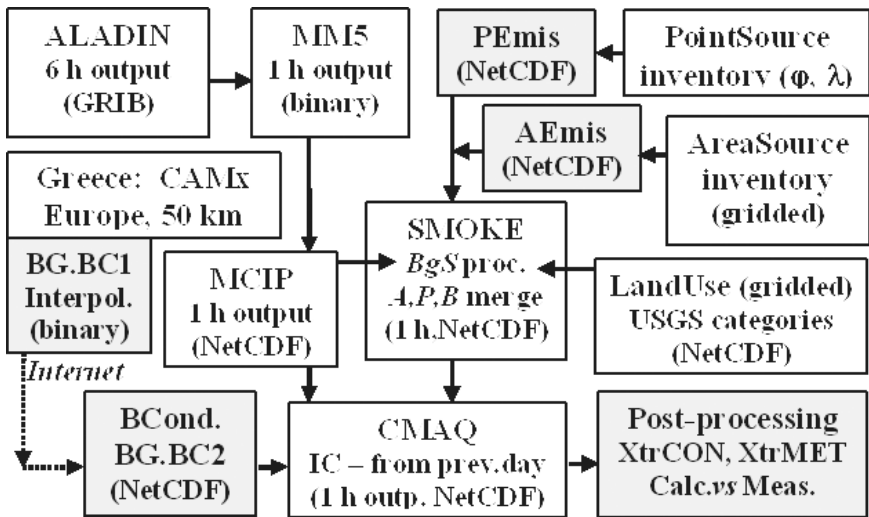


Fig. 4. Data flow of calculations for estimating climate change impact on air quality

The 1-day data flow of this System is shown in Fig. 4. The white boxes present Models-3 compounds; the brown (dark grey) boxes denote specialized FORTRAN interface programs and the green (light grey) boxes present the databases - input to the System. First of these databases is the ALADIN data that drives MM5. MM5s output feeds MCIP (Meteorology-Chemistry Interface Processor of CMAQ) that produces formatted meteorological input to both CMAQ and SMOKE. The RegCM3/CAMx database [5, 6] is based in the Aristotle University of Thessaloniki, Greece, and contains air pollution calculations for Europe on a 50-km grid. It is used for off-line interpolations to the CMAQ boundary points (module BG.BC1), results uploaded to a dedicated server in Sofia. This data is used by the System for on-line elaboration of CMAQ boundary conditions (module BG.BC2).

CMAQ demands its emission input in specific format reflecting the time evolution of all pollutants accounted for by the chemical mechanism used. The emissions, produced by different sources (anthropogenic and natural) must be united. The main groups of emission sources are Area Sources (AS), Large Point Sources (LPS) and Biogenic Sources (BgS). Their processing is described in details in [11, 12, 13]. In spite the investigated periods are quite long, the emission scenario in CECILIA is fixed only EMEP 50 km emission database (<http://webdab.emep.int/>) for 2000 is applied. In fact, its disaggregation to higher resolution (1515 km) made by [15] is exploited. The use of only year 2000 emissions is to estimate only the influence of meteorology changes on pollution levels in regional and local aspect.

As shown in Fig. 4, the AS- and LPS-inventories feed the interface programs AEmis and PEmis which produce respective emission files. The gridded land-use information for the domain is introduced in SMOKE which calculates the biogenic emissions exploiting the ambient meteorological data provided by MCIP. Finally, SMOKE merges these files as to produce common emission input for CMAQ.

The post-processing programs XtrCON and XtrMET extract part of the pollutants and meteorological fields for archiving. Only surface values of 17 most important pollutants are saved on hourly basis.

Before starting the assessment of climate change impact on air quality, validation of model results is needed in order to assure the quality and credibility of delivered output. There are quite few measurements in the territory of Bulgaria able to be used. Fortunately, it occurs that during all the year 2000 in the frame of research project hourly ozone measurements have been performed in two points in Bulgaria - peak Rojen, Rhodopi mountain, and Ahtopol, a small town at Black sea coast ([3]). This data is the main source for testing model results certainty.

The scatter diagrams of calculated vs. measured hourly ozone concentrations are displayed in Fig. 5. One can notice that almost all scatter points are into FA2 boundaries. The fitting lines are quite close to the ideal fitting line; the correlation coefficients of both fits are quite high that reflects the good quality of the simulation. These and other comparisons point out that the System can be used with some certainty.

5 Intensive Long-Term Simulations Using GRID Computation

For CECILIA purpose, long-term simulations for three 10-year periods: Control Run (CR: 1991-2000), Near Future (NF: 2041-2050) and Far Future (FF: 2091-2100) are performed. The respective archives consist of daily files, produced by XtrCON. A number of additional programs were created for fast retrieving this data and calculating various statistics.

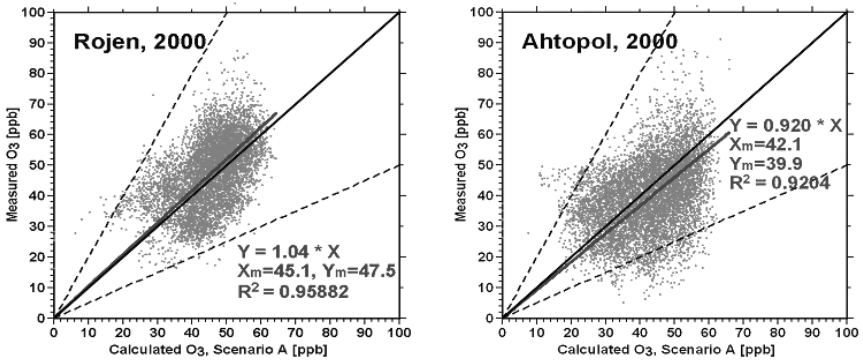


Fig. 5. Tendency in the yearly accumulated precipitation in %: $100*(NF-CR)/CR$ (left) and $100*(FF-CR)/CR$ (right)

Because of the enormous volume of calculations GRID technology is used for this application. GRID provides access to huge computational and storage resources. In addition, GRID provides a failover capability even if some cluster is down temporarily, there will be other resources available for completing the work.

The application consists of several scripts, where the output of one script feeds the next one. These dependencies have been successfully wrapped-up in one master script that is submitted to the GRID as a job and thus completing the computation for the whole period consists in submitting a series of consecutive GRID jobs. The jobs require MPI support and the main part of the computational time is taken by the MM5 and CMAQ computations. In both cases a locally verified version of mpich2 is used and deployed on-the-fly as part of the job. The total size of the input data for one Grid job is approximately 3.5 GB, when initial and boundary conditions for one-month simulation are included. The generated output is less than 1 GB. The application is fully gridified, except the post-processing and analysis of the data, which is done at a local workstation. The application benefited from the experience and the advices of scientists from the Institute of Parallel Processing with Bulgarian Academy of Sciences (IPP-BAS), who have already developed some impressive GRID applications like SALUTE [1]. In this work mostly the resources from IPP-BAS provided within the framework of the EC FP7 SEE-GRID-SCI project (<http://www.see-grid-sci.eu/>) were used, where the applications acronym is CCIAQ (Climate Change Impact on Air Quality).

6 Contemporary Pollution Levels and Its Changes due to Climate Ones

Here, results only for ozone are given. For other pollutants see [10]. In Fig. 6, left, the now-a-day climatic ozone field is presented compared with the field of 10-year mean Averaged Daily Maxima (ADM). One can notice some resemblance

between space patterns but the values are quite different. The ADM are usually preferred because the high ozone values are harmful most of all. In Fig. 7, the Ozone ADM differences are displayed. It is seen that the differences "FF-CR" are higher than "NF-CR", being in correspondence with temperature tendencies (see Fig. 2). The maximal values are over 3 ppb that is about 5-10% of ADM maximums. There is a definite separation between positive and negative difference values in both cases they are positive in the plain parts of the domain and negative in the mountain areas. The maximal positive differences are concentrated around the main pollution sources.

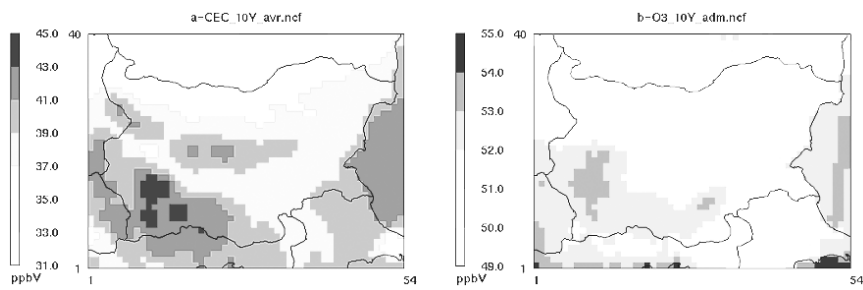


Fig. 6. 10-year mean Ozone concentration (left) and ADM (right), CR

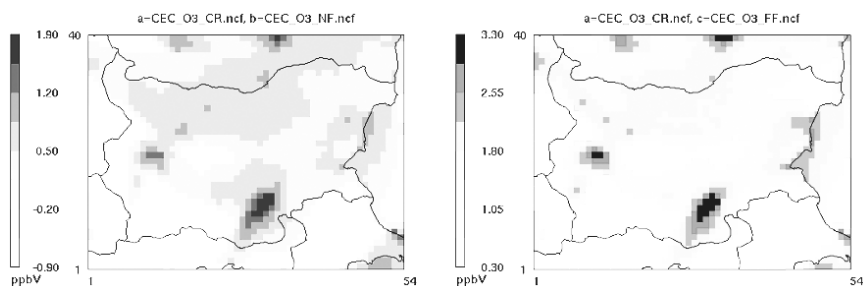


Fig. 7. Differences in the 10-year ADM fields for "NF-CR" (left) and "FF-NF" (right)

7 Conclusion

In the present work some estimates of the dimensions of the climate changes for the region of Bulgaria are presented. They are calculated by the ALADIN-Climate model following IPCC scenario A1B. This Scenario is considered as the most realistic one. Considering the changes in the air pollution levels (at least for the 4 main pollutants), it can be concluded that the changes are quite small (5-10% of maximal values). The space distribution of these changes also shows definite specifics. The maximal positive changes are located around the main pollution sources in the region (as well outside the region but closed to the

boundaries and in such a way reflected in the boundary conditions). In the main part of the country and in higher extent in the mountain areas the changes are at decreasing pollution levels. As to the order of magnitude of maximal positive changes, they are highest in FF-CR case. This behavior of the climatic changes of air pollution levels reflects to a certain extent the behavior of the changes in the meteorological climatic values presented.

Acknowledgments

This study is made under the financial support of European Commission FP6 Integrated Project CECILIA and the FP7 projects SEE-GRID-SCI. The presented results were not possible without the experience obtained during the participation in the FP5 project BULAIR, FP6 Network of Excellence ACCENT and the FP6 Integrated Project QUANTIFY. The contacts within the framework of the NATO SfP Project No. 981382 were extremely simulating as well. Deep gratitude is due to all organizations providing free of charge data and software used in this study, namely US EPA, US NCEP and European institutions like EMEP, EEA and many others. Special thanks to the Netherlands Organization for Applied Scientific Research (TNO) for providing with the high-resolution European anthropogenic emission inventory and emission time allocation profiles. Without the databases and the software produced by these organizations the present study would not be possible.

References

1. Atanasov, E., Gurov, T., Karaivanova, A.: SALUTE application for Quantum Transport - New Grid Implementation Scheme. In: Proceedings of the Spanish Conference on e-Science Grid Computing, March 1-2, pp. 23-32 (2007)
2. Deque, M., Piedelievre, J.-P.: High-Resolution climate simulation over Europe. *Climate Dynamics* 11, 321-339 (1995)
3. Donev, E., Zeller, K., Avramov, A.: Preliminary background ozone concentrations in the mountain and coastal areas of Bulgaria. *Environmental Pollution* 17, 281-286 (2002); IPCC: IPCC Special report: Emissions scenarios (2000), ISBN: 92-9169-113-5, http://www.grida.no/publications/other/ipcc_tar/?src=/climate/ipcc_tar/wg3/081.htm
4. Katragkou, E., Zanis, P., Tegoulas, I., Melas, D., Krueger, B.C., Huszar, P., Halenka, T.: Tropospheric Ozone over Europe: An air quality model evaluation for the period 1990-2001. In: Proceedings of IX EMTE National-International Conference of Meteorology-Climatology and Atmospheric Physics, Thessaloniki, Greece, May 28-31, p. 649 (2008)
5. Krueger, B.C., Katragkou, E., Tegoulas, I., Zanis, P., Melas, D., Coppola, E., Rauscher, S.A., Huszar, P., Halenka, T.: Regional decadal photochemical model calculations for Europe concerning ozone levels in a changing climate. *Idjrs - Quarterly Journal of the Hungarian Meteorological Service* 112(3-4), 285-300 (2008)
6. Somot, S., Spiridonov, V., Marquet, P., Deque, M.: Climate version of the LAM ALADIN. In: Workshop on Regional Climate Modeling, MAGMA EC Project No EVG3-CT-2002-80006, Prague (2004)

7. Spiridonov, V., Braun, A., Deque, M., Somot, S.: High resolution climate adaptation of ERA40 data over the Bulgarian domain. In: Workshop on Regional Climate Modeling, MAGMA EC Project No EVG3-CT-2002-80006, Prague (2004)
8. Spiridonov, Deque, M., Somot, S.: ALADIN-CLIMATE: from the origins to present date. ALADIN Newsletter 29 (2005)
9. Spiridonov, V., Syrakov, D., Ganev, K., Prodanova, M., Bogachev, A., Miloshev, N., Jordanov, G., Slavov, K.: Model estimates of regional climate changes and its impact on the air quality over Bulgaria. In: 19th International Symposium ECOLOGY & SAFETY, Sunny Beach, Bulgaria, vol. 4, Part 1, June 7 - 11 (2010) (on a CD), ISSN: 1313-2563, <http://www.science-journals.eu>
10. Syrakov, D., Ganev, K., Spiridonov, V., Prodanova, M., Bogatchev, A., Miloshev, N., Jordanov, G.: Assessment of climate change impact on air pollution levels in Bulgaria. In: 7th International Conference on Air Quality Science and Application Istanbul, March 24-27 (2009a) (on a CD)
11. Syrakov, D., Prodanova, M., Miloshev, N., Ganev, K., Jordanov, G., Spiridonov, V., Bogatchev, A., Katragkou, E., Melas, D., Poupkou, A., Markakis, K.: Climate Change Impact Assessment of Air Pollution Levels in Bulgaria. In: Lirkov, I., Margenov, S., Wasniewski, J. (eds.) LSSC 2009. LNCS, vol. 5910, pp. 538-545. Springer, Heidelberg (2010)
12. Syrakov, D., Ganev, K., Prodanova, M., Slavov, K., Miloshev, N., Jordanov, G.: Impact of climate change on air pollution levels in Bulgaria: Contemporary climate. In: 18th International Symposium ECOLOGY & SAFETY, Sunny Beach, Bulgaria, June 8 - 12, vol. 3, Part 1, pp. 22-31 (2009c) (on a CD) ISSN: 1313-2563, <http://www.science-journals.eu>
13. Uppala, S.M., Kllberg, P.W., Simmons, A.J., Andrae, U., da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hlm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., McNally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J.: The ERA-40 re-analysis. *Quart. J. R. Meteorol. Soc.* 131, 2961-3012 (2005), doi:10.1256/qj.04.176
14. Visschedijk, A.J.H., Zandveld, P.Y.J., Denier van der Gon, H.A.C.: A High Resolution Gridded European Emission Database for the EU Integrate Project GEMS, TNO-report 2007-A-R0233/B, Apeldoorn, The Netherlands (2007)

Genetic Algorithms Based Parameter Identification of Yeast Fed-Batch Cultivation

Maria Angelova, Stoyan Tzonkov, and Tania Pencheva

Institute of Biophysics and Biomedical Engineering - Bulgarian Academy of Sciences
105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
`maria.angelova@clbme.bas.bg`, `tzonkov@clbme.bas.bg`,
`tania.pencheva@clbme.bas.bg`

Abstract. Different kinds of genetic algorithms have been investigated for a parameter identification of a fermentation process. Altogether eight realizations of genetic algorithms have been presented - four of simple genetic algorithms and four of multi-population ones. Each of them is characterized with a different sequence of implementation of main genetic operators, namely selection, crossover and mutation. A comparison of considered eight kinds of genetic algorithms is presented for a parameter identification of a fed-batch cultivation of *S. cerevisiae*. All kinds of multi-population algorithms lead to considerable improvement of the optimization criterion value but for more computational time. Among the considered multi-population algorithms, the best one has an operators' sequence of crossover, mutation and selection. Different kinds of considered simple genetic algorithms lead to similar values of the optimization criterion but the genetic algorithm with an operators' sequence of mutation, crossover and selection is significantly faster than the others.

1 Introduction

Fermentation processes (FP) are widely used in different branches of industry, i.e. in the production of pharmaceuticals, chemicals and enzymes, yeast, foods and beverages. Live microorganisms play an important role in these processes so their peculiarities predetermine some specific characteristics of FP as modeling and control objects. As complex, nonlinear, dynamic systems with interdependence and time-varying process variables, FP are a serious challenge for modelling and further high-quality control. An important step for adequate modeling of nonlinear models of FP is the choice of a certain optimization procedure for model parameter identification. The conventional optimization methods can not overcome the limitations of FP, while genetic algorithms (GA), as stochastic global optimization method, are quite promising. GA are a direct random search technique for finding global optimal solution in complex multidimensional search space. GA have a lot of advantages such as hard problems solving, noise tolerance, easy to interface and hybridize. All these properties predetermine GA as suitable and more workable for the optimization of highly non-linear problems, especially for a parameter identification of fermentation process models [1, 2, 8–10].

Simple (SGA) and multi-population (MpGA) genetic algorithm, as presented initially in Goldberg [5] search a global optimal solution using three main genetic operators in a sequence selection, crossover and mutation. For the purpose of this investigation, SGA and MpGA with such sequence are denoted respectively as SGA-SCM and MpGA-SCM. Many improved variations of the SGA and MpGA have been developed [1, 4, 7, 10]. Among them are the modified genetic algorithm [10] with a sequence crossover, mutation and selection, here denoted as SGA-CMS, and consequent modification of MpGA based on such exchange [1], here denoted as MpGA-CMS. In these algorithms selection operator has been processed after performing of crossover and mutation. The main idea for such operators' sequence is to prevent the loss of reached good solution by either crossover or mutation or both operators. SGA-CMS applied to a parameter identification of *E. coli* fed-batch cultivation [10] and further tested also for a parameter identification of *S. cerevisiae* fed-batch cultivation [1] improves the optimization capability of the algorithm, decreasing decision time. MpGA-CMS applied to a parameter identification of *S. cerevisiae* fed-batch cultivation decreases the algorithm calculation time and improves significantly the decision adequacy compared to SGA. Obtained promising results applying SGA-CMS and MpGA-CMS encourage more investigations to be performed in order further improvements of the algorithms to be found.

In GA as presented initially in Holland [6] and further in Goldberg [5], the operator mutation is usually applied after the operator crossover. The basic idea of GA is to imitate the mechanics of natural selection and genetics, so one can make an analogy with the processes occurring in the nature. As such, the probability mutation to come first and then crossover is comparable to the idea both processes to occur in a reverse order. The purpose of this study is to investigate the influence of the genetic operators' sequence selection, crossover and mutation in SGA and MpGA. Presented in [5] SGA-SCM and MpGA-SCM, as well as developed SGA-CMS [10] and MpGA-CMS [1] have been compared with four new proposed kinds with exchanged sequence of mutation and crossover operators. Obtained altogether eight kinds of the SGA and the MpGA have been compared in terms of accuracy and performance for a parameter identification of *S. cerevisiae* fed-batch cultivation.

2 Implementation of Exchanged Operators' Sequence of Crossover and Mutation in Simple and Multi-population Genetic Algorithms

The ideology of implementation of GA for the parameter identification purposes could be summarized as follows. The chromosomes represents the models parameters and corresponding objective function value is associated to each chromosome. The objective function is used to provide a measure of how individuals have performed in the problem domain. In the case of minimization problem, the

fitted individuals will have the lowest numerical value of the associated objective function. This raw measure of fitness is only used as an intermediate stage in determining the relative performance of individuals in genetic algorithms. The selection algorithm chooses individuals for reproduction on the basis of their relative fitness. Selected chromosomes, through reproduction, crossover and mutation, form a new population. Generated in that way population is used for a further run of the algorithm. The GA is terminated when a certain number of generations is fulfilled, a mean deviation in the population is satisfied, or when a particular point in the search space is encountered.

Simple genetic algorithm (denoted here as SGA-SCM) guides the mechanism of evaluation implementing the three main operators in a sequence selection, crossover and mutation. Presently four modifications of SGA and MpGA are elaborated and demonstrated, implementing the exchange of operators' sequence crossover and mutation. Newly presented modifications are as follows:

- SGA-SMC - a modification of the developed in [5] SGA-SCM;
- SGA-MCS - a modification of the developed in [10] SGA-CMS;
- MpGA-SMC - a modification of the developed in [5] MpGA-SCM;
- MpGA-MCS - a modification of the developed in [1] MpGA-CMS.

Since the MpGA are more complex than SGA, and as a case with most exchanges towards the originally presented by Goldberg GA, the elaboration of MpGA-MCS is shortly presented below. Multi-population genetic algorithm is a single population genetic algorithm, in which many populations, called subpopulations, evolve independently from each other for a certain number of generations. After a certain number of generations (isolation time), a number of individuals are distributed between the subpopulations. In the beginning, the MpGA generates a random population of n chromosomes, i.e. suitable solutions for the problem. In order to prevent the loss of reached good solution by either crossover or mutation or both operators, selection operator has been processed after performing of crossover and mutation [1]. The new modification presented here is that, the individuals are reproduced processing firstly mutation, followed by crossover. The elements of chromosome are a bit changed when a newly created offspring mutates, after that the genes from parents combine to form a whole new chromosome during the crossover. After the reproduction, the MpGA-MCS calculates the fitness values for the offspring and the best fitted individuals are selected to replace the parents. Then the algorithm evaluates the objective values (cost values) of the individuals in the current population and according to that the new chromosome is created. the MpGA is terminated when a certain number of generations is fulfilled.

Proposed exchange in a operators' sequence mutation and crossover has been also applied towards SGA-SCM, SGA-CMS and MpGA-SCM. This results in new algorithm modifications considered in this investigation and denoted as SGA-SMC, SGA-MCS and MpGA-SMC respectively.

3 Parameter Identification of *S. cerevisiae* Fed-Batch Cultivation Using Different Kinds of Simple and Multi-population Genetic Algorithms

Experimental data of *S. cerevisiae* fed-batch cultivation is obtained in the Institute of Technical Chemistry - University of Hannover, Germany. The cultivation of the yeast *S. cerevisiae* is performed in a 2 l reactor, using a Schatzmann medium [8]. The initial liquid volume is 1.3 l. Glucose in feeding solution is 35 g/l. The temperature was controlled at 30C, the pH at 5.5. The stirrer speed was set to 1200 rpm. The aeration rate was kept at 300 l/h. Biomass and ethanol were measured off-line, while substrate (glucose) and dissolved oxygen were measured on-line.

Mathematical model of *S. cerevisiae* fed-batch cultivation is commonly described as follows, according to the mass balance [11]:

$$\frac{dX}{dt} = \mu X - \frac{F}{V} X \quad (1)$$

$$\frac{dS}{dt} = -q_S X + \frac{F}{V} (S_{in} - S) \quad (2)$$

$$\frac{dE}{dt} = q_E X - \frac{F}{V} E \quad (3)$$

$$\frac{dO_2}{dt} = -q_{O_2} X + k_L^{O_2} a (O_2^* - O_2) \quad (4)$$

$$\frac{dV}{dt} = F \quad (5)$$

where X is the concentration of biomass, [g/l]; S - concentration of substrate (glucose), [g/l]; E - concentration of ethanol, [g/l]; O_2 - concentration of oxygen, [%]; O_2^* - dissolved oxygen saturation concentration, [%]; F - feeding rate, [l/h]; V - volume of bioreactor, [l]; $k_L^{O_2} a$ - volumetric oxygen transfer coefficient, [1/h]; S_{in} - initial glucose concentration in the feeding solution, [g/l]; μ , q_S , q_E , q_{O_2} - specific growth/utilization rates of biomass, substrate, ethanol and dissolved oxygen, [1/h].

Considered here fed-batch cultivation of *S. cerevisiae* is characterized with keeping glucose concentration equal or below to its critical level ($S_{crit} = 0.05$ g/l) and with sufficient dissolved oxygen in the broth $O_2 \geq O_{2crit}$ ($O_{2crit} = 18\%$). This state corresponds to so called mixed oxidative state according to functional state modeling approach [11]. As presented in [11], the specific growth rate is generally found to be a sum of two terms, one describing the contribution of sugar and the other - the contribution of ethanol to yeast growth. Both terms have the structure of Monod model. Monod model is also used for the specific ethanol and sugar consumption rates. Dissolved oxygen consumption rate is obtained as a sum of two terms, which are directly proportional to the specific glucose rate

and specific ethanol production rate, respectively. Hence, specific rates in Eqs. (1)-(5) are presented as follows:

$$\begin{aligned}\mu &= \mu_{2S} \frac{S}{S + k_S} + \mu_{2E} \frac{E}{E + k_E}, q_S = \frac{\mu_{2S}}{Y_{SX}} \frac{S}{S + k_S}, \\ q_E &= -\frac{\mu_{2E}}{Y_{EX}} \frac{E}{E + k_E}, q_{O_2} = q_E Y_{OE} + q_S Y_{OS}\end{aligned}\quad (6)$$

where μ_{2S} , μ_{2E} - maximum growth rates of substrate and ethanol, [1/h]; k_S , k_E - saturation constants of substrate and ethanol, [g/l]; Y_{ij} - yield coefficients, [g/g].

As an optimization criterion, mean square deviation between the model output and the experimental data obtained during cultivation has been used:

$$J_Y = \sum (Y - Y^*)^2 \rightarrow \min \quad (7)$$

where Y and Y^* are the experimental and model predicted data respectively, $Y = [X, S, E, O_2]$.

Parameter identification of the model (1)-(5) has been performed using *Genetic Algorithm Toolbox* in *Matlab 5.3* environment [3]. All the computations are performed using a PC Intel Pentium 4 (2.4 GHz) platform running Windows XP. Consequently eight kinds of genetic algorithms - four kinds of SGA and four kinds of MpGA, four of them newly presented here, have been applied for the purpose of a parameter identification of *S. cerevisiae* fed-batch cultivation. A comparison between performances of four kinds of SGA is presented in Table 1, while Table 2 presents results obtained using four kinds of MpGA.

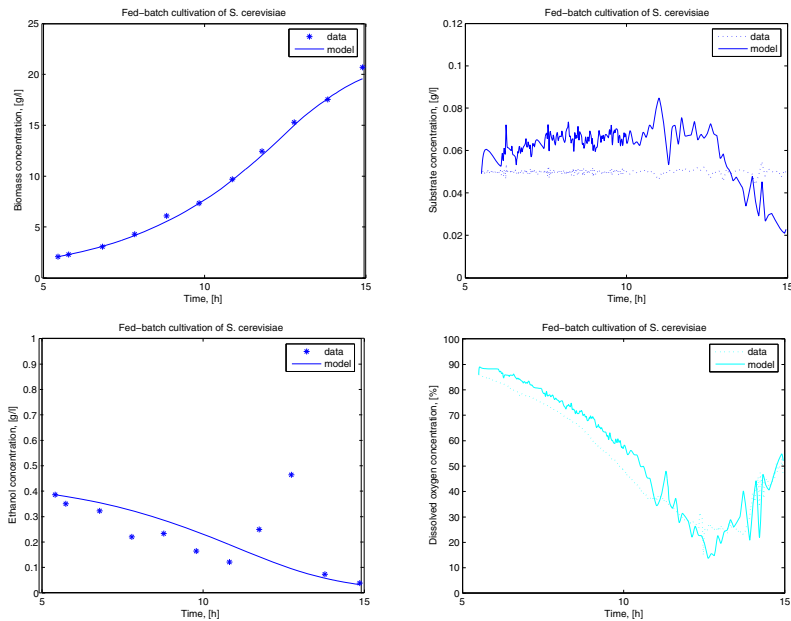
Table 1. Results from model parameter identification using different kinds of SGA

Parameter	SGA-SCM	SGA-SMC	SGA-CMS	SGA-MCS
J_Y	0.0223	0.0221	0.0225	0.0223
CPU time, s	73.8281	73.4688	64.8281	59.5156
μ_{2S} , 1/h	0.9616	0.9038	0.9211	0.9119
μ_{2E} , 1/h	0.0971	0.1320	0.0872	0.0966
k_S , g/l	0.1154	0.1119	0.1176	0.1109
k_E , g/l	0.7963	0.7990	0.7620	0.7987
Y_{SX} , g/g	0.4279	0.4072	0.4279	0.4316
Y_{EX} , g/g	1.2898	1.7699	1.2898	1.3170
$k_L^{O_2} a$, 1/h	38.5895	116.4160	127.2898	141.1076
Y_{OS} , g/g	313.8285	898.6292	989.8014	993.2537
Y_{OE} , g/g	234.7797	281.1797	62.6547	166.6377

As shown in Table 1, the optimization criterion values obtained with four types of standard genetic algorithms are very similar. Hopefully, there is no loss of adequacy of the model when the operator mutation is performed before crossover. Moreover, proposed modification in the algorithm reduces time of

Table 2. Results from model parameter identification using different kinds of MpGA

Parameter	MpGA-SCM	MpGA-SMC	MpGA-CMS	MpGA-MCS
J_Y	0.0144	0.0145	0.0144	0.0145
CPU time, s	100.6563	98.0625	95.6094	100.4688
μ_{2S} , 1/h	0.9000	0.9012	0.9003	0.9073
μ_{2E} , 1/h	0.1447	0.0967	0.1342	0.0549
k_S , g/l	0.1500	0.1499	0.1500	0.1500
k_E , g/l	0.8000	0.7739	0.8000	0.7647
Y_{SX} , g/g	0.3944	0.4131	0.4076	0.4271
Y_{EX} , g/g	6.9156	4.8402	6.5616	2.7389
$k_L^{O_2} a$, 1/h	101.6394	71.5478	95.7177	98.3150
Y_{OS} , g/g	808.7495	569.6776	753.0205	772.7289
Y_{OE} , g/g	522.0352	759.6290	282.2053	449.7269

**Fig. 1.** Experimental data and model prediction for biomass, substrate, ethanol and dissolved oxygen concentrations

reaching of a global minimum. While the implementation of SGA-SMC compared to SGA-SCM does not lead to significant decrease ($<1\%$) of decision time, the use of SGA-MCS reduces the time with 9% compared to SGA-CMS. The fastest algorithm SGA-MCS achieves the global minimum for 24% less time than SGA-SCM. Presented here comparison shows that the implementation of the operators in a sequence of mutation, crossover and then selection is the most optimal in attitude of rate with reserved high adequacy of the decision.

As it is seen from Table 2, the values of the optimization criterion obtained using multi-population genetic algorithms are also comparable. The expected improvement in CPU time of MpGA-SMC towards MpGA-SCM has been observed (with about 3%), while MpGA-MCS reaches the decision slowly than MpGA-CMS. Hence, the fastest MpGA has an operators sequence of crossover, mutation and selection and reaches the decision with 5% faster than MpGA-SCM.

The results presented in Table 1 for SGA and those in Table 2 for MpGA have been compared too. The value of the optimization criterion in MpGAs is about 50% less than the criterion in SGA. Unfortunately, MpGA need more time to reach the global minimum. That is why it is up to the user to make a decision which type of GA to use as a compromise between the time consumption and model precision.

Due to the similarity of the results from the implementation of all considered here eight types of GA, only these obtained with MpGA-CMS (as the fastest and the most precise among the MpGA) are here presented. Fig. 1 presents results from experimental data and model prediction respectively for biomass, substrate, ethanol and dissolved oxygen.

4 Analysis and Conclusions

In this investigation altogether four modifications two of SGA and two of MpGA have been proposed, implementing the exchanged operators sequence of mutation and crossover. Newly suggested SGA-SMC, SGA-MCS, MpGA-SMC and MpGA-MCS have been developed and compared respectively to SGA-SCM, SGA-CMS, MpGA-SCM and MpGA-CMS for the purposes of a parameter identification of a fed-batch cultivation of *S. cerevisiae*. Implementation of the main genetic operators in order mutation, crossover and selection in SGA significantly improves calculation time of the algorithm without affecting to the model adequacy. SGA-MCS solves the optimization problem 9% faster than SGA-CMS and 24% than SGA-SCM. Four kinds of MpGA lead to significant improvement of about 50% of the optimization criterion value but for more computational time. Among the considered MpGA, the fastest and the most precise one implements an operators sequence of crossover, mutation and selection. Finally, comparing SGA and MpGA it is up to the user to make a decision which type of GA to use as a compromise between the time consumption and model precision.

Acknowledgements

This work is partially supported by the European Social Fund and Bulgarian Ministry of Education, Youth and Science under Operative Program "Human Resources Development", grant BG051PO001-3.3.04/40 and National Science Fund of Bulgaria, grant number DID 02-29 "Modeling Processes with Fixed Development Rules".

References

1. Angelova, M., Tzonkov, S., Pencheva, T.: Modified multi-population genetic algorithm for yeast fed-batch cultivation parameter identification. *Int. J. Bioautomation* 13(4), 163–172 (2009)
2. Carrillo-Ureta, G.E., Roberts, P.D., Becerra, V.M.: Genetic algorithms for optimal control of beer fermentation. In: *Proc. of the 2001 IEEE Int. Symp. on Intelligent Control*, Mexico City, Mexico, pp. 391–396 (2001)
3. Chipperfield, A.J., Fleming, P., Pohlheim, H., Fonseca, C.M.: Genetic algorithm toolbox for use with MATLAB. Users guide, version 1.2. Dept. of Automatic Control and System Engineering, University of Sheffield, UK (1994)
4. Cordon, O., Herrera, F.: Hybridizing genetic algorithms with sharing scheme and evolution strategies for designing approximate fuzzy rule-based systems. *Fuzzy Sets and Systems* 118, 235–255 (2001)
5. Goldberg, D.: Genetic algorithms in search, optimization and machine learning. Addison-Wiley Publishing Company, Massachusetts (1989)
6. Holland, J.: Adaptation in natural and artificial systems. MIT Press, Cambridge (1975)
7. Kuo, R.J., Chen, C.H., Hwang, Y.C.: An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. *Fuzzy Sets and Systems* 118, 21–45 (2001)
8. Pencheva, T., Roeva, O., Hristozov, I.: Functional state approach to fermentation processes modelling. In: Tzonkov, S., Hitzmann, B. (eds.) *Prof. Marin Drinov. Academic Publishing House*, Sofia (2006)
9. Ranganath, M., Renganathan, S., Gokulnath, C.: Identification of bioprocesses using genetic algorithm. *Bioprocess Engineering* 21, 123–127 (1999)
10. Roeva, O.: A modified genetic algorithm for a parameter identification of fermentation processes. *Biotechnol. and Biotechnol. Equip.* 20, 202–209 (2006)
11. Zhang, X.-C., Visala, A., Halme, A., Linco, P.: Functional state modelling approach for bioprocesses: local models for aerobic yeast growth processes. *J. Proc. Contr.* 4(3), 127–134 (1994)

Intuitionistic Fuzzy Interpretations of Conway's Game of Life

Lilija Atanassova¹ and Krassimir Atanassov²

¹IICT – Bulgarian Academy of Sciences, Acad. G. Bonchev str. bl.2,
1113 Sofia, Bulgaria

l.c.atanassova@gmail.com

²CLBME – Bulgarian Academy of Science, Acad. G. Bonchev str, bl 105,
1113 Sofia, Bulgaria
krat@bas.bg

In memory of our teacher in informatics Prof. Peter Barnev.

Abstract. Conway's Game of Life is a popular heuristic zero-player game, devised by John Horton Conway in 1970, and it is the best-known example of a cellular automaton. Its "universe" is an infinite two-dimensional orthogonal grid of square cells, each of which is in one of two possible states, alive or dead. Every cell interacts with its eight neighbours, which are the cells that are directly horizontally, vertically, or diagonally adjacent. In a stepwise manner, the state of each cell in the grid preserves or alternates with respect to a given list of rules. Intuitionistic fuzzy sets (IFS) are an extension of Zadeh's fuzzy sets, which introduce a degree of membership and a degree of non-membership whose sum is equal to or less than 1 and the complement to 1 is called a degree of uncertainty. The article proposes an intuitionistic fuzzy estimation of the cells' state in a modified Game of Life. For each cell we can define its IF estimation as a pair consisting of the degrees l_p and l_a , namely degrees of presence and absence of life, where $l_p + l_a \leq 1$. In the classical Conway's Game of Life, the alive and dead states correspond to the elementary IF estimations $\langle 1, 0 \rangle$ and $\langle 0, 1 \rangle$. The article presents the formulas for calculating the IF state of liveliness of each cell, as functions of the current states of the cell's neighbours. Criteria of liveliness will be also determined in terms of IFS.

1 Introduction

Conway's Game of Life is devised by John Horton Conway in 1970, and already 40 years it is an object of research, software implementations and modifications. In [1] there is a list of many papers devoted to Conway's Game of Life. In 1976, the authors who were then students in Sofia University, also introduced one modification of this game. In the present paper another modification will be introduced. It is based on the idea of the intuitionistic fuzziness.

The standard Conway's Game of Life has a "universe" which is an infinite two-dimensional orthogonal grid of square cells, each of which is in one of two

possible states, alive or dead, or as we learned the game from the lectures of Prof. Barnev in the middle of 1970s, in the square there is an asterisk or not.

Every cell interacts with its eight neighbours, namely the cells that are directly adjacent either in horizontal, vertical, or diagonal direction. In a stepwise manner, the state of each cell in the grid preserves or alternates with respect to a given list of rules.

Here, we will discuss some versions of the game in which we will keep the condition for the necessary number of existing neighbours asterisks for birth or dying of an asterisk in some square. For our aims, we will use elements of intuitionistic fuzzy set theory (see, [2]).

2 Remarks on Intuitionistic Fuzzy Sets and Logic

Let us have some set of propositions S . To every proposition p from this set there are assigned real numbers $\mu(p)$ and $\nu(p)$, such that $\mu(p), \nu(p) \in [0, 1]$ and

$$\mu(p) + \nu(p) \leq 1.$$

These numbers correspond to the “truth degree” and to the “falsity degree” of p .

Let this assignment be provided by an evaluation function V , defined over a set of propositions S in such a way that:

$$V(p) = \langle \mu(p), \nu(p) \rangle.$$

Everywhere below, we shall assume that for the two variables x and y the equalities $V(p) = \langle a, b \rangle$, $V(q) = \langle c, d \rangle$ ($a, b, c, d, a + b, c + d \in [0, 1]$) hold.

Obviously, when V is an ordinary fuzzy truth-value estimation, then

$$b = 1 - a.$$

For the needs of the discussion below, we shall define the notion of Intuitionistic Fuzzy Tautology (IFT, see, [2]) by:

$$p \text{ is an IFT if and only if } a \geq b,$$

while p will be a tautology iff $a = 1$ and $b = 0$.

In some definitions we shall use functions sg and $\overline{\text{sg}}$:

$$\text{sg}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \quad \overline{\text{sg}}(x) = \begin{cases} 0 & \text{if } x > 0 \\ 1 & \text{if } x \leq 0 \end{cases}$$

When values $V(p)$ and $V(q)$ of the propositional forms p and q are known, the evaluation function V can be extended also for the operations “conjunction” (two forms), “disjunction” (two forms), “implication” (about 140 different forms), “negation” (34 different forms) and others. Here, we will only need operation “negation” ($\neg$), that for proposition p will have the forms given in Table 1.

Table 1.

$\neg_1$	$\langle x, b, a \rangle$
$\neg_2$	$\langle x, \overline{\text{sg}}(a), \text{sg}(a) \rangle$
$\neg_3$	$\langle x, b, a.b + a^2 \rangle$
$\neg_4$	$\langle x, b, 1 - b \rangle$
$\neg_5$	$\langle x, \overline{\text{sg}}(1 - b), \text{sg}(1 - b) \rangle$
$\neg_6$	$\langle x, \overline{\text{sg}}(1 - b), \text{sg}(a) \rangle$
$\neg_7$	$\langle x, \overline{\text{sg}}(1 - b), a \rangle$
$\neg_8$	$\langle x, 1 - a, a \rangle$
$\neg_9$	$\langle x, \overline{\text{sg}}(a), a \rangle$
$\neg_{10}$	$\langle x, \overline{\text{sg}}(1 - b), 1 - b \rangle$
$\neg_{11}$	$\langle x, \text{sg}(b), \overline{\text{sg}}(b) \rangle$
$\neg_{12}$	$\langle x, b.(b + a), a.(b^2 + a + b.a) \rangle$
$\neg_{13}$	$\langle x, \text{sg}(1 - a), \overline{\text{sg}}(1 - a) \rangle$
$\neg_{14}$	$\langle x, \text{sg}(b), \overline{\text{sg}}(1 - a) \rangle$
$\neg_{15}$	$\langle x, \overline{\text{sg}}(1 - b), \overline{\text{sg}}(1 - a) \rangle$
$\neg_{16}$	$\langle x, \overline{\text{sg}}(a), \overline{\text{sg}}(1 - a) \rangle$
$\neg_{17}$	$\langle x, \overline{\text{sg}}(1 - b), \overline{\text{sg}}(b) \rangle$
$\neg_{18}$	$\langle x, b.\text{sg}(a), a.\text{sg}(b) \rangle$
$\neg_{19}$	$\langle x, b.\text{sg}(a), 0 \rangle$
$\neg_{20}$	$\langle x, b, 0 \rangle$
$\neg_{21}$	$\langle x, \min(1 - a, \text{sg}(a)), \min(a, \text{sg}(1 - a)) \rangle$
$\neg_{22}$	$\langle x, \min(1 - a, \text{sg}(a)), 0 \rangle$
$\neg_{23}$	$\langle x, 1 - a, 0 \rangle$
$\neg_{24}$	$\langle x, \min(b, \text{sg}(1 - b)), \min(1 - b, \text{sg}(b)) \rangle$
$\neg_{25}$	$\langle x, \min(b, \text{sg}(1 - b)), 0 \rangle$
$\neg_{26}$	$\langle x, b, a.b + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{27}$	$\langle x, 1 - a, a.(1 - a) + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{28}$	$\langle x, b, (1 - b).b + \overline{\text{sg}}(b) \rangle$
$\neg_{29}$	$\langle x, a.b + \overline{\text{sg}}(1 - b), a.(a.b + \overline{\text{sg}}(1 - b)) + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{30}$	$\langle x, a.b, a.(a.b + \overline{\text{sg}}(1 - b)) + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{31}$	$\langle x, (1 - a).a + \overline{\text{sg}}(a), a.((1 - a).a + \overline{\text{sg}}(a)) + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{32}$	$\langle x, (1 - a).a, a.((1 - a).a + \overline{\text{sg}}(a)) + \overline{\text{sg}}(1 - a) \rangle$
$\neg_{33}$	$\langle x, b.(1 - b) + \overline{\text{sg}}(1 - b), (1 - b).(b.(1 - b) + \overline{\text{sg}}(1 - b)) + \overline{\text{sg}}(b) \rangle$
$\neg_{34}$	$\langle x, b.(1 - b), (1 - b).(b.(1 - b) + \overline{\text{sg}}(1 - b)) + \overline{\text{sg}}(b) \rangle$

3 Intuitionistic Fuzzy Criteria of Existence, Birth and Death of an Asterisk

Let us have a plane tessellated with squares. Let in some of these squares there be symbols “*”, meaning that the squares are “alive”. Now we will extend this construction of the Game of Life to some new forms.

Let us assume that the square $\langle i, j \rangle$ is assigned a pair of real numbers $\langle \mu_{i,j}, \nu_{i,j} \rangle$, so that $\mu_{i,j} + \nu_{i,j} \leq 1$. We can call the numbers $\mu_{i,j}$ and $\nu_{i,j}$ degree of existence and degree of non-existence of symbol “*” in square $\langle i, j \rangle$. Therefore, $\pi(i, j) = 1 - \mu_{i,j} - \nu_{i,j} \leq 1$ will correspond to the degree of uncertainty, e.g., lack of information about existence of an asterisk in the respective square.

Below we will formulate a series of different criteria for correctness of the intuitionistic fuzzy interpretations that will include as a particular case the standard game.

3.1 Six Criteria of Existence of an Asterisk

We will suppose that there exists an asterisk in square $\langle i, j \rangle$ if:

- (1.1) $\mu_{i,j} > 0.5$. Therefore $\nu_{i,j} < 0.5$. In the particular case, when $\mu_{i,j} = 1 > 0.5$ we obtain $\nu_{i,j} = 0 < 0.5$, i.e., the standard existence of the asterisk.
- (1.2) $\mu_{i,j} \geq 0.5$. Therefore $\nu_{i,j} \leq 0.5$. Obviously, if case (1.1) is valid, then case (1.2) also will be valid.
- (1.3) $\mu_{i,j} > \nu_{i,j}$. Obviously, case (1.1) is particular case of the present one, but case (1.2) is not included in the currently discussed case for $\mu_{i,j} = 0.5 = \nu_{i,j}$.
- (1.4) $\mu_{i,j} \geq \nu_{i,j}$. Obviously, cases (1.1), (1.2) and (1.3) are particular cases of the present one.
- (1.5) $\mu_{i,j} > 0$. Obviously, cases (1.1), (1.2) and (1.3) are particular cases of the present one, but case (1.4) is not included in the currently discussed case for $\mu_{i,j} = 0.0 = \nu_{i,j}$.
- (1.6) $\nu_{i,j} < 1$. Obviously, cases (1.1), (1.2) and (1.3) are particular cases of the present one, but case (1.5) is not included in the currently discussed case for $\mu_{i,j} = 0.0$.

From these criteria it follows that if one is valid – let it be the s -th criterion ($1 \leq s \leq 6$) then we can assert that the asterisk exists with respect to the s -th criterion and, therefore, it will exist with respect to all other criteria, whose validity follows from the validity of the s -th criterion.

On the other hand, if s -th criterion is not valid, then we will say that the asterisk does not exist with respect to s -th criterion. It is very important that in this case the square may not be absolutely empty. It is appropriate to tell that the square $\langle i, j \rangle$ is totally empty, if its degrees of existence and non-existence are $\langle 0, 1 \rangle$.

It is suitable to tell that the square is s -full if it contains an asterisk with respect to the s -th criterion and that the same square is s -empty if it does not satisfy the s -th criterion.

For the aims of the game-method for modelling, it will be suitable to use (with respect to the type of the concrete model) one of the first four criteria for existence of an asterisk. Let us say for each fixed square $\langle i, j \rangle$ that therein is an asterisk by s -th criterion for $1 \leq s \leq 4$, if this criterion confirms the existence of an asterisk.

3.2 Four Criteria for the Birth of an Asterisk

In the standard game, the rule for birth of a new asterisk is: the (empty) square has exactly 2 or 3 neighbouring squares containing asterisks. Now we will formulate a series of different rules that will include as a particular case the standard rule.

- **2.1** (extended standard rule): The s -empty square has exactly 2 or 3 neighbouring s -full squares. Obviously, this rule for birth is a direct extension of the standard rule.
- **2.2** (pessimistic rule): For the natural number $s \geq 2$, the s -empty square has exactly 2 or 3 neighbouring $(s - 1)$ -full squares.
- **2.3** (optimistic rule): For the natural number $s \leq 5$, the s -empty square has exactly 2 or 3 neighbouring $(s + 1)$ -full squares.
- **2.4** (average rule): Let $M_{i,j}$ and $N_{i,j}$ be, respectively, the sums of the μ -degrees and of the ν -degrees of all neighbours of the s -empty square. Then the inequality

$$\frac{1}{4} \cdot N_{i,j} \leq M_{i,j} \leq \frac{3}{8} \cdot N_{i,j}$$

holds.

3.3 Four Criteria for the Death of an Asterisk

In the standard game the rule for the death of an existing asterisk is: the (full) square has exactly 2 or 3 neighborhood squares containing asterisks. Now we will formulate a series of different rules that will include as a particular case the standard rule.

- **3.1** (extended standard rule): The s -full square has less than 2 or more than 3 neighboring s -full squares. Obviously, this rule for dying is a direct extension of the standard rule.
- **3.2** (pessimistic rule): For the natural number $s \geq 2$, the s -full square has less than 2 or more than 3 neighboring $(s - 1)$ -full squares.
- **3.3** (optimistic rule): For the natural number $s \leq 5$, the s -full square has less than 2 or more than 3 neighboring $(s + 1)$ -full squares.
- **3.4** (average rule): Let $M_{i,j}$ and $N_{i,j}$ be, respectively, the sums of the μ -degrees and of the ν -degrees of all neighbours of the s -full square. Then one of the inequalities

$$\frac{1}{4} \cdot N_{i,j} > M_{i,j} \quad \text{or} \quad M_{i,j} > \frac{3}{8} \cdot N_{i,j}$$

holds.

4 Intuitionistic Fuzzy Rules for Changing of the Game-Field

In the standard game the game-field is changed by the above mentioned rules for birth and death of the asterisks. Now, we will discuss some intuitionistic fuzzy

rules for changing of the game-field. They use the separate forms of operation “negation”.

Let us suppose that in a fixed square there is an asterisk if and only if the square is s -full. Therefore, we tell that in the square there is no asterisk if and only if the square is not s -full. In this case we can call this square s -empty.

As we saw above, the difference between standard and intuitionistic fuzzy form of the game is the existence of values corresponding to the separate squares. In the standard case they are 1 or 0, or “there exists an asterisk”, “there is no asterisk”. In the intuitionistic fuzzy form of the game we have pairs of real numbers as in the case when the asterisk exists, as well as in the opposite case. In the classical case, the change of the status of the square is obvious. In the intuitionistic fuzzy we can construct different rules. They are of two types.

The first type contains two modifications of the standard rule:

- **4.1** (extended standard rule): If an s -full square $\langle i, j \rangle$ must be changed, then we can use negation $\neg_1$ for pair $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and in a result we will obtain pair $\langle \nu_{i,j}, \mu_{i,j} \rangle$.
- **4.2** (non-standard, or intuitionistic fuzzy rule): If an s -full square $\langle i, j \rangle$ must be changed, then we can use any of the other negations $\neg_m$ from Table 1 ($2 \leq m \leq 34$).

The second type contains three non-standard modifications. The standard rule and the above two rules for changing of the current content of the fixed square (existence or absence of an asterisk) are related only to this content. Now, we can include a new parameter, that conditionally can be called “*the influence of the environment*”.

- **5.1** (optimistic (s, m)-rule) If an s -full/empty square $\langle i, j \rangle$ must be changed, then we can use m -th negation $\neg_m$ to pair (before change) $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and to juxtapose to it the pair $\langle \mu_{i,j}^*, \nu_{i,j}^* \rangle$, so that

$$\mu_{i,j}^* = \max(\mu'_{i,j}, \max_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \mu_{u,v})$$

$$\nu_{i,j}^* = \min(\nu'_{i,j}, \min_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \nu_{u,v}),$$

where

$$\langle \mu'_{i,j}, \nu'_{i,j} \rangle = \neg_m \langle \mu_{i,j}, \nu_{i,j} \rangle$$

and $\max^*, \min^*$ mean that we use only values that are connected to s -empty/full squares.

- **5.2** (optimistic-average (s, m)-rule) If an s -full/empty square $\langle i, j \rangle$ must be changed, then we can use m -th negation $\neg_m$ to pair (before change) $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and to juxtapose to it the pair $\langle \mu_{i,j}^*, \nu_{i,j}^* \rangle$, so that

$$\mu_{i,j}^* = \max(\mu'_{i,j}, \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \mu_{u,v})$$

$$\nu_{i,j}^* = \min(\nu'_{i,j}, \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \nu_{u,v}),$$

where $\langle \mu'_{i,j}, \nu'_{i,j} \rangle$ is as above, $\sum^*$ mean that we use only values that are connected to s -empty/full squares and $t(i, j)$ is the number of these squares.

- **5.3** (average (s, m) -rule) If an s -full/empty square $\langle i, j \rangle$ must be changed, then we can use m -th negation $\neg_m$ to pair (before change) $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and to juxtapose to it the pair $\langle \mu_{i,j}^*, \nu_{i,j}^* \rangle$, so that

$$\mu_{i,j}^* = \frac{1}{2}(\mu'_{i,j} + \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \mu_{u,v})$$

$$\nu_{i,j}^* = \frac{1}{2}(\nu'_{i,j} + \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \nu_{u,v}),$$

where $\langle \mu'_{i,j}, \nu'_{i,j} \rangle$, $\sum^*$ and $t(i, j)$ are as in 5.1 and 5.2.

- **5.4** (pessimistic-average (s, m) -rule) If an s -full/empty square $\langle i, j \rangle$ must be changed, then we can use m -th negation $\neg_m$ to pair (before change) $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and to juxtapose to it the pair $\langle \mu_{i,j}^*, \nu_{i,j}^* \rangle$, so that

$$\mu_{i,j}^* = \min(\mu'_{i,j}, \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \mu_{u,v})$$

$$\nu_{i,j}^* = \max(\nu'_{i,j}, \frac{1}{t(i,j)} \sum_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \nu_{u,v}),$$

where $\langle \mu'_{i,j}, \nu'_{i,j} \rangle$, $\sum^*$ and $t(i, j)$ are as in 5.1 and 5.2.

- **5.5** (pessimistic (s, m) -rule) If an s -full/empty square $\langle i, j \rangle$ must be changed, then we can use m -th negation $\neg_m$ to pair (before change) $\langle \mu_{i,j}, \nu_{i,j} \rangle$ and to juxtapose to it the pair $\langle \mu_{i,j}^*, \nu_{i,j}^* \rangle$, so that

$$\mu_{i,j}^* = \min(\mu'_{i,j}, \min_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \mu_{u,v})$$

$$\nu_{i,j}^* = \max(\nu'_{i,j}, \max_{u \in \{i-1, i, i+1\}; v \in \{j-1, j, j+1\}; \langle u, v \rangle \neq \langle i, j \rangle}^* \nu_{u,v}),$$

where $\langle \mu'_{i,j}, \nu'_{i,j} \rangle$ and $\max^*, \min^*$ are as in case 5.1.

5 Conclusion

Here a series of modifications of the laws of the Conway's Game of Life functioning, based on intuitionistic fuzzy set theory, were introduced for the first time. In the next authors' research new modifications of this game will be described. We will continue in two directions.

First, we will modify the standard game using other elements of the intuitionistic fuzzy set theory, e.g, the modal, topological and level operators, defined in it.

Second: we will modify the rules of the game, as we already prepared this in our previous research, e.g. [3–5].

Acknowledgements

The authors are grateful for the support provided by the projects DID-02-29 “Modelling processes with fixed development rules” and BIn-2/09 “Design and development of intuitionistic fuzzy logic tools in information technologies” funded by the National Science Fund, Bulgarian Ministry of Education, Youth and Science.

References

1. Conway's Game of Life. In Wikipedia, The Free Encyclopedia (May 8, 2010), http://en.wikipedia.org/w/index.php?title=Conway's_Game_of_Life&oldid=360850256
2. Atanassov, K.: Intuitionistic Fuzzy Sets. Springer, Heidelberg (1999)
3. Atanassov, K., Atanassova, L.: A game method for modelling. In: Antonov, L. (ed.) Third International School, Automation and Scientific Instrumentation, Varna, pp. 229–232 (1984)
4. Atanassov, K.: On a combinatorial game-method for modelling. In: Advances in Modelling & Analysis, vol. 19(2), pp. 41–47. AMSE Press (1994)
5. Atanassov, K., Atanassova, L., Sassellov, D.: On the combinatorial game-method for modelling in astronomy. *Comptes Rendus de l'Academie bulgare des Sciences*, Tome 47(9), 5–7 (1994)

Ant Colony Optimization Approach to Tokens' Movement within Generalized Nets

Vassia Atanassova¹ and Krassimir Atanassov²

¹ IICT – Bulgarian Academy of Sciences, Acad. G. Bonchev str. bl.2,
1113 Sofia, Bulgaria
vassia.atanassova@gmail.com

² CLBME – Bulgarian Academy of Science, Acad. G. Bonchev str, bl 105,
1113 Sofia, Bulgaria
krat@bas.bg

Abstract. Generalized Nets (GNs) is a concept, extending the concept of Petri Nets and the rest of its modifications: an apparatus for modelling of parallel and concurrent processes. GNs have been applied to modelling of processes in the field of artificial intelligence, and in particular to metaheuristic methods for solving of optimizational problems, like the transportation problem, the travelling salesman problem, the knapsack problem, etc. An important venue of application of GNs is the area of Ant Colony Optimization (ACO). So far, GNs have been used as a method for description of the ACO procedures. The present article for the first time adopts the opposite approach: it discusses the possibility for optimization of the GN tokens' movement, using ACO algorithms.

Keywords: Ant colony optimization, Generalized net, Modelling.

1 Introduction

Generalized Nets (GNs, see [1,2]) is a concept extending the concept of Petri nets and the rest of its modifications. One of the aspects of generalization is the fact that the GN transitions possess an index matrix of predicates, determining the conditions for tokens' transfer from any input place of the transition to any output place. On the other hand, the tokens enter the GN with their initial characteristics and during their transfer from the input to the output places of the transition, they are assigned new characteristics by means of special characteristic functions.

GNs have been applied to modelling of processes in the field of artificial intelligence (expert systems, neural networks, pattern recognition, machine learning, etc.), and in particular to metaheuristic methods for solving of optimizational problems like the transportation problem, the travelling salesman problem, the knapsack problem. An important venue of application of GN is the area of Ant Colony Optimization (ACO, see [4,5,7]). So far, GN have been used as a method for description of the ACO procedures.

The present article for the first time adopts the opposite approach: it discusses the possibility for optimization of the GN tokens' movement, using ACO algorithms.

2 Short Remarks on GN Theory

Broadly speaking, the GN is a bipartite directed graph consisting of a set of vertices called transitions and another set of vertices called places. Both of them reflect the static nature, or the infrastructure, of the modeled process, while its dynamic nature is represented by a set of tokens initialized with certain starting characteristic which move from the input to the output places of the model having their characteristics changed. In other words, the tokens are instances of the modeled process, or individuals, who keep track of their history. The tokens, their characteristic functions and the conditions for transfer (as coded in index matrices of the transitions' predicates) reflect the logic of the modeled process.

The formal definition of the GN requires firstly the definition of the net's building block, namely the transition.

$$TR = \langle P_{IN}, P_{OUT}, time, dur, IMP, IM_C, bool \rangle$$

where

- P_{IN} is the finite nonempty set of the input places (obligatory).
- P_{OUT} is the finite nonempty set of the output places (obligatory).
- $time$ is the current time of the transition's activation (optional).
- dur is the current duration of the transition's active state (optional).
- IMP is the index matrix of predicates, determining the conditions for tokens' transfer through the transition (obligatory).
- IM_C is the matrix, determining the number of tokens that may transfer from the i -th input to the j -th output of the transition (optional).
- $bool$ is the Boolean type of the transition (optional).

On this basis, the formal definition of the GN is given, comprising of four sets of components demonstrating respectively the static, dynamic, temporal nature of the net and its memory:

$$GN = \langle \langle TRS, \pi_{TR}, \pi_P, c, f, \theta_{ACT}, \theta_{DUR} \rangle, \\ \langle TKN, \pi_{TKN}, \theta_{TKN} \rangle, \\ \langle T, t^0, t^* \rangle, \\ \langle X_{INIT}, X_{NEW}, n \rangle \rangle$$

Static components of the net

- TRS - set of transitions (obligatory);
- π_{TR} - function, giving the priorities of the transitions (optional);
- π_P - function, giving the priorities of the places (optional);
- c - function, giving the capacities of the places (optional);
- f - function that evaluates the degree of the predicates in IMP (it may be restricted to the $\{false, true\}$ -set, or in the $[0; 1]$ interval or in the multiset $[0; 1] \times [0; 1]$ (optional);
- θ_{ACT} - function, giving the next time moment when a given transition would get activated; the value is calculated only when the transition has stopped being active (optional);
- θ_{DUR} - function, giving the duration of the active state (optional).

Dynamic components of the net

- TKN - set of tokens (obligatory);
- π_{TKN} - function, giving the tokens' priorities (optional);
- θ_{TKN} - function, giving the time moment when a given token will enter the net (optional).

Temporal components of the model, defined according to a global time scale

- T - the moment of time when the net would start functioning (optional);
- t^o - the elementary incremental step of the time scale (optional);
- t^* - the total duration of net's functioning (optional).

Characteristic components (memory) of the net

- X_{INIT} - the set of initial characteristics that tokens acquire on entering the net (obligatory);
- X_{NEW} - the characteristic function, which assigns new characteristics of the tokens on their transferring via given transition (obligatory);
- n - function, giving the maximal number of characteristics for storing in a token's memory (optional):
 - $n = 0$ - the token stores no characteristics in its memory;
 - $n = 1$ - the token stores only its current characteristic;
 - $n = k$ - the token stores only the last k acquires characteristics;
 - $n = \infty$ - all token's characteristics are stored in the memory.

Different operations, relations and operators are defined over the transitions of the GNs and over the same nets. A variety of different types of GN-extensions are defined and each of them is proved [1,2] to be a conservative extension of the ordinary GNs.

Now, we will give the general algorithm for tokens transfer in the frames of a transition at time moment $t_1 = TIME$ (the current GN time-moment), as described in [2]. In the following section we will present our idea for modifying some of its steps, which is inspired by the ACO.

(A01) Sort the input and output places of the transitions by their priorities. The tokens from a given input place are divided into two groups. The first one contains those tokens that can be transferred to the transition output, the second contains the rest (the motivation for this will be clear from the next steps of the algorithm). Let the two parts be denoted by " $P_1(l)$ " and " $P_2(l)$ ", respectively, where l is the corresponding place.

(A02) Sort the tokens from group P_1 of the input places (following the order from A01) by their priorities. Let the index matrix R correspond to the index matrix IM_P . Thus, the (u, v) -th element of R is

$$R_{u,v} = \begin{cases} 1, & \text{if the } (u, v)\text{-th predicate } r_{u,v} \text{ is true} \\ 0, & \text{if the } (u, v)\text{-th predicate } r_{u,v} \text{ is false or if the value is} \\ & \text{determined by A03.} \end{cases}$$

(A03) Assign a value 0 to all elements of R for which either

(a) the input place which corresponds to the respective predicate is empty (the part P_1 is empty); or

(b) the output place which corresponds to the respective predicate is full; or

(c) the current capacity of the arc between the corresponding input and output places is 0.

(A04) Calculate the values of the other elements of IM_P and assign the obtained values to the elements of R .

(A05) Calculate the values of the characteristic functions related to the corresponding output places in which tokens will enter. Assign these characteristics to the entering tokens.

(A06) Perform the following for each input place by the order of input place priorities:

a) select the tokens with the highest priority in this input place;

b) transfer the selected tokens to all output places, for which the corresponding predicate enables this (the tokens go to group P_2 of the output places).

(A07) Transfer the tokens with the highest priority, for which all calculated values of the predicates are equal to “*false*” to the group P_2 of the corresponding places. In this group, also transfer all tokens that cannot be transferred to the corresponding output places because these places have already been filled with tokens from other places with higher priorities.

(A08) Add t^0 to the current time, i.e., $TIME := TIME + t^0$.

(A09) Check whether the value of the current time is less than $t_1 + t_2$ (the time-components of the considered transition).

(A10) If the answer to the question in A09 is “yes”, go to A02 (to update the tokens' order in the places).

(A11) If the answer to the question in A09 is “no”, terminate the current functioning of the transition.

3 Main Results

Up to now, GNs have been used for modelling, simulation, in certain cases management, optimization or machine learning of real processes. For example, there has been developed a GN model that makes decisions of the structure of a neural network that solves particular problems with predefined accuracy of the solution and duration of functioning [3]. However, as of today, no GN has been constructed in a way to optimize models that take place inside of it. An idea of such a GN is the Self-Modifying GN, but up to now no such net has been constructed and published. Now, using ideas from the ACO algorithm we will initiate the first step towards researching the possibility for construction of a particular GN that is capable of taking decisions for changes in some of its own parameters.

In other words, the basic idea of this work is to combine the notions of GNs and ACO in the opposite way of those utilized so far. As of today, the concept of GNs was used to describe different variants of the ACO algorithm [6]. Here

we follow the reverse approach, applying the principle of ants' movement to the tokens' movement throughout the net. To do so, we have to pay attention to the following considerations and interpretations of the elements of the ACO algorithm in terms of GNs.

- The ACO algorithm can be reduced to finding optimal paths through graphs. Hence, here we will utilize the fact that the GN has a graphic structure that may be interpreted as a graph.
- The artificial ants are interpreted as the tokens in the GN.
- The pheromone trails are used by the artificial ants in the ACO algorithm as communication medium: once the agents have found a solution they depose these traces, i.e. communicate their discovery with the agents-to-come. In terms of GNs, this information shall be given the form of a list of the net's places that have been visited.

The changes in the pheromone's intensity (increase due multiple ants using the track, or decrease due to evaporation) are modeled by changes in the characteristics of some appropriately chosen tokens. These changes will be an object of discussion in a next authors' research.

Let us have a GN that models a concrete process, of which we know:

- the separate stages as represented by the net's transitions,
- the carriers of dynamic behaviour, as represented by individual tokens, and
- the moments of the tokens' entering the GN.

If we possess all of this information about the process, we will be capable of constructing an adequate GN model of this process, while if a part of this information is missing, our GN model will not be complete but partial. Below, we will discuss how we may approach to replacing some of the missing data. We will show how we can generate appropriate values of some of the model's parameters, which will be derived by the modeled process itself, making the assumption that it functions in an optimal way.

For instance, one case of incomplete information of the modeled process is to assume that in the real process we miss the data about the durations of the transitions from one state to another, as well as the durations of the separate states. Another possible situation (when we happen to have more information) is if we know the durations of the separate sub-processes, but we do not know what characteristics we may assign to the net's tokens that describe the dynamics of the process. It is an even more interesting case when we possess part of this information, as well. For each of these three examples we may design a GN that reflects the relations between the separate parts of the modeled process. It is a priori clear that at least this knowledge ought to be in being.

The present article will deal only with the first of the so described scenarios.

Let us take a GN with 1 or more input places and 1 or more output places. Let us make the following assumptions:

- On each step every transition of the net is fired (gets activated) and its active state continues 1 time unit.

- All tokens are allowed to split.
- The tokens' memory is unlimited, i.e. all tokens may store an indefinite number of characteristics.
- Each token have the initial characteristic of the moment of time when it enters the GN.

In order to describe the first example we shall assume that the capacities of the places are equal to infinity. In this case, every token transfer from the input place to each of the output places of the respective transition. It is sufficient in this case to have exactly one token entering each input place, because otherwise the next-to-come will repeat the exact ways of splitting and the routes of the preceding tokens.

In each place, the tokens obtain as a new characteristic the place's identifier (the current place's identifier is added to the list of identifiers of all previous, already attended, places in the net).

The so-described GN precisely copycats the idea of an ACO procedure with a finite number of ants, each of which is here represented by a GN token. The token, which starts from the i -th input place and is the first to reach the j -th output place of the net, will possess as characteristics the shortest route between both of these places.

When describing the second example, we will have to assume that the capacities of the places are finite numbers, in particular 1. In this case, we are able to take into account the eventual instances of route clogging, and for this reason this case is more interesting than the first one. Now, we can have a new token entering each input place only when the previous token had already left the place.

In each place the tokens obtain as a current characteristic the place's identifier as well as the moment of entering. In the end, the final token characteristic will also include the calculated total time of token's movement throughout the net.

It is appropriate to have the process of tracking the tokens' movement described in the GN itself, i.e. to have the net self-controlling. For this purpose, we add to the given GN a new transition T (see Fig. 1) with only one place P that serves both for input and output place. Only one token α loops in the transition. The transitions T , the place P and the token α are assigned the minimal possible priorities among their likes. In this way, on each step of the net's functioning we provide for the token α to make its move after all other tokens in the net, and allow it to obtain as a current characteristic the current distribution of tokens per places.



Fig. 1.

After the end of the net's functioning, we will determine the shortest route with respect to either time, or length by:

- tracing the routes of the individual tokens,
- determining the lengths of the paths, and
- rendering account of the time spent by the tokens in the net (Case 2).

Behind the so constructed GN construction, another important aspect can be perceived, namely the criteria of intended optimization. Our experience with the classical ACO has led us to the understanding that it is the time of taking the route and the length of the path in the GN, as generated by the GN structure, that are most important criteria for optimization. Now it is clear that this statement is valid for the first of the discussed cases, but it is invalid in the second case, when the duration and the length of the path may be fully independent criteria and the optimization may be conducted per both of them, in parallel.

On the basis of the accumulated information, we may built a simulation model in which the tokens transfer from input to output places with probabilities corresponding to the profits laid on the respective routes.

Now we will discuss the possible applications of the so constructed GN.

As we already mentioned, there is a point in using it only in cases when we possess incomplete information of the modeled process. In the first case discussed above, we may complicate the research by determining the lengths of the paths from the i -th place, which is not an input place of the GN, up to the j -th output place, and then we can apply the following procedure:

1. For each (say, t -th) transition, we determine the number of the output places, via which a token that has started from the i -th place which is an input place for this transition, will reach the j -th place which is output place for the whole net. Let this transition possess s_t output places and let their route lengths be, respectively, $p_1^t, p_2^t, \dots, p_{s_t}^t$. Then we determine the number $a_t = \sum_{i=k}^{s_t} \frac{1}{p_k^t}$.
2. We determine the numbers $\alpha_k^t = \frac{p_k^t}{a_t}$ ($1 \leq k \leq s_t$).
3. The predicate of the index matrix of transition t that corresponds to the fixed i -th place and the k -th output place be $P_{i,k} = "r \in \left[\sum_{u=1}^{k-1} \frac{1}{p_u^t}, \sum_{u=1}^k \frac{1}{p_u^t} \right]"$, where r is a random number in the $[0, 1]$ interval.

Following this procedure, the token from the i -th place will advance to an output place with a probability that corresponds to the length of the route to the j -th output place of the net. Moreover, the shorter the path, the larger the probability for the token to move towards this very place. This ensures the optimal movement of the nets around the net.

In contrast with the first case, in the second case we assume that tokens enter the net in every time moment. Now, for t -th transition and for its k -th output place ($1 \leq k \leq s_t$) we will obtain that the tokens (whose number is q_k), which have passed through it, will arrive in the net's j -th output place for time periods of $Q_{k,1}^t, Q_{k,2}^t, \dots, Q_{k,q_k}^t$. These time periods can be different, because in the second case the tokens can spend time waiting in some places. All of these tokens will travel a path of length p_k^t (as in the first case). Now, we can determine the

average duration for tokens' transfer: $D_k^t = \frac{1}{q_k} \sum_{l=1}^{q_k} Q_{k,l}^t$. By analogy with the first case, we can determine the numbers $\beta_k^t = \sum_{k=1}^{s_t} \frac{1}{D_k^t}$, that we can use instead of α_k^t , constructed above.

4 Conclusion

This paper contains the general idea and the first step towards optimization of the GN functioning by the ant colony optimization algorithm. A next authors' research will be especially devoted to the formal description and exploration of the rest two cases, as well as other situations that may occur in the GNs. It must be noted that using the above discussed ideas a self-organizing GN can be constructed, which makes references to one of the open problems in artificial intelligence, namely the problem with self-reference and self-modifying algorithms (see [8,9]).

Acknowledgments

This work has been supported by the Bulgarian National Science Fund under grants No. DID-02-29 "Modelling Processes with Fixed Development Rules" and DTK-02-44 "Effective Monte Carlo Methods for Large-Scale Scientific Problems".

References

1. Atanassov, K.: Generalized Nets. World Scientific, Singapore (1991)
2. Atanassov, K.: On Generalized Nets Theory. Prof. M. Drinov Publishing House, Sofia (2007)
3. Atanassov, K., Sotirov, S.: Optimization of a Neural Network of Self-organizing Maps Type with Time-Limits by a Generalized Net. *Advanced Studies on Contemporary Mathematics* 13(2), 213–220 (2006)
4. Dorigo, M., Gambardella, L.M.: Ant Colony system: A Cooperative Learning Approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation* 1, 53–66 (1997)
5. Dorigo, M., Stutzle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
6. Fidanova, S., Atanassov, K.: Generalized Net Models of the Process of Ant Colony Optimization. *Issues in Intuitionistic Fuzzy Sets and Generalized Nets* 7, 108–114 (2008)
7. Fidanova, S., Marinov, P.: Intuitionistic fuzzy estimation of the ant methodology. *Int. J. of Cybernetics and Information Technology* 9(2), 79–88 (2009)
8. Marshall, J., Hofstadter, D.: Beyond Copycat: Incorporating Self-Watching into a Computer Model of High-Level Perception and Analogy-Making. In: Gasser, M. (ed.) *Online Proceedings of the 1996 Midwest Artificial Intelligence and Cognitive Science Conference*, Indiana University, Bloomington (1996)
9. Turney, P.: (2007),
<http://apperceptual.wordpress.com/2007/12/18/open-problems/>

Start Strategies of ACO Applied on Subset Problems

Stefka Fidanova¹, Krassimir Atanassov², and Pencho Marinov¹

¹ IPP – Bulgarian Academy of Sciences, Acad. G. Bonchev str. bl.25A,
1113 Sofia, Bulgaria

{stefka,pencho}@parallel.bas.bg

² CLBME – Bulgarian Academy of Science, Acad. G. Bonchev str, bl 105,
1113 Sofia, Bulgaria
krat@bas.bg

Abstract. Ant Colony Optimization is a stochastic search method that mimic the social behavior of real ants colonies, which manage to establish the shortest routs to feeding sources and back. Such algorithms have been developed to arrive at near-optimum solutions to large-scale optimization problems, for which traditional mathematical techniques may fail. In this paper on each iteration estimations of the start nodes of the ants are made. Several start strategies are prepared and combined. Benchmark comparisons among the strategies are presented in terms of quality of the results. Based on this comparison analysis, the performance of the algorithm is discussed along with some guidelines for determining the best strategy. The study presents ideas that should be beneficial to both practitioners and researchers involved in solving optimization problems.

1 Introduction

The difficulties associated with using mathematical optimization on large-scale engineering problems, have contributed to the development of alternative solutions. Linear programming and dynamic programming techniques, for example, often fail in solving NP-hard problems with large number of variables. To overcome these problems, researchers have proposed metaheuristic methods for searching near-optimal solutions to problems. One of the most successful metaheuristic is Ant Colony Optimization (ACO).

Real ants foraging for food lay down quantities of pheromone (chemical cues) marking the path that they follow. An isolated ant moves essentially at random but an ant encountering a previously laid pheromone will detect it and decide to follow it with high probability and thereby reinforce it with a further quantity of pheromone. The repetition of the above mechanism represents the auto-catalytic behavior of a real ant colony where the more the ants follow a trail, the more attractive that trail becomes.

ACO is inspired by real ant behavior to solve hard combinatorial optimization problems. Examples of hard optimization problems are Traveling Salesman Problem [9], Vehicle Routing [10], Minimum Spanning Tree [7], Constrain Satisfaction [5], Knapsack Problem [3,4], etc. The ACO algorithm uses a colony of

artificial ants that behave as cooperative agents in a mathematical space where they are allowed to search and reinforce pathways (solutions) in aiming to find the optimal ones. The problem is represented by graph and the ants walk on the graph to construct solutions. The solutions are represented by paths in the graph. After the initialization of the pheromone trails, the ants construct feasible solutions, starting from random nodes, and then the pheromone trails are updated. At each step the ants compute a set of feasible moves and select the best one (according to some probabilistic rules) to continue the rest of the tour. The structure of the ACO algorithm is shown by the pseudocode below. The transition probability $p_{i,j}$, to choose the node j when the current node is i , is based on the heuristic information $\eta_{i,j}$ and the pheromone trail level $\tau_{i,j}$ of the move, where $i, j = 1, \dots, n$.

$$p_{i,j} = \frac{\tau_{i,j}^a \eta_{i,j}^b}{\sum_{k \in Unused} \tau_{i,k}^a \eta_{i,k}^b}, \quad (1)$$

where *Unused* is the set of unused nodes of the graph.

The higher the value of the pheromone and the heuristic information, the more profitable it is to select this move and resume the search. In the beginning, the initial pheromone level is set to a small positive constant value τ_0 ; later, the ants update this value after completing the construction stage. ACO algorithms adopt different criteria to update the pheromone level.

Ant Colony Optimization

```

Initialize number of ants;
Initialize the ACO parameters;
while not end-condition do
    for k=0 to number of ants
        ant k choses start node;
        while solution is not constructed do
            ant k selects higher probability node;
        end while
    end for
    Update-pheromone-trails;
end while
    
```

Fig. 1. Pseudocode for ACO

The pheromone trail update rule is given by:

$$\tau_{i,j} \leftarrow \rho \tau_{i,j} + \Delta \tau_{i,j}, \quad (2)$$

where ρ models evaporation in the nature and $\Delta \tau_{i,j}$ is new added pheromone which is proportional to the quality of the solution.

The novelty in this work is the use of estimations of start nodes with respect to the quality of the solution and thus to better manage the search process.

Various start strategies and their combinations are offered. The problem used for testing is Multiple Knapsack Problem (MKP) like a representative of subset problems.

The rest of the paper is organized as follows. In section 2 estimation of start regions and several start strategies are proposed. In section 3 the new ideas are applied on MKP and computational results are analyzed. At the end some conclusions and directions for future work are done.

2 Start Strategies

The known ACO algorithms create a solution starting from random node. But for some problems, especially subset problems, it is important from which node the search process starts. For example if an ant starts from a node which does not belong to the optimal solution, the probability to construct it is zero. In this paper is offered several start strategies. The aim is to use the experience of the ants from previous iteration to choose the better starting node. Other authors use this experience only by the pheromone, when the ants construct the solutions.

Let the graph of the problem has m nodes. The set of nodes is divided on N subsets. There are different ways for dividing. Normally, the nodes of the graph are randomly enumerated. An example for creating of the subsets, without lost of generality, is: the node number one is in the first subset, the node number two - in the second subset, etc., the node number N is in the N -th subset, the node number $N + 1$ is in the first subset, etc. Thus the number of nodes in the separate subsets are almost equal. After the first iteration the estimations $D_j(i)$ and $E_j(i)$ are introduced of the node subsets, where $i \geq 2$ is the number of the current iteration. $D_j(i)$ is an estimation how good is the subset j and $E_j(i)$ is an estimation how bad is the subset j . $D_j(i)$ and $E_j(i)$ are weight coefficients of j -th node subset ($1 \leq j \leq N$), which are calculated by the following formulas:

$$D_j(i) = \frac{i \cdot D_j(i-1) + F_j(i)}{i}, \quad (3)$$

$$E_j(i) = \frac{i \cdot E_j(i-1) + G_j(i)}{i}, \quad (4)$$

where $i \geq 2$ is the current iteration and for each j ($1 \leq j \leq N$):

$$F_j(i) = \begin{cases} \frac{f_{j,A}}{n_j} & \text{if } n_j \neq 0 \\ F_j(i-1) & \text{otherwise} \end{cases}, \quad (5)$$

$$G_j(i) = \begin{cases} \frac{g_{j,B}}{n_j} & \text{if } n_j \neq 0 \\ G_j(i-1) & \text{otherwise} \end{cases}, \quad (6)$$

and $f_{j,A}$ is the number of the solutions among the best $A\%$, and $g_{j,B}$ is the number of the solutions among the worst $B\%$, where $A + B \leq 100$, $i \geq 2$ and $\sum_{j=1}^N n_j = n$, where n_j ($1 \leq j \leq N$) is the number of solutions obtained by ants starting from nodes subset j . Initial values of the weight coefficients are: $D_j(1) = 1$ and $E_j(1) = 0$.

Let threshold E for $E_j(i)$ and D for $D_j(i)$ be fixed, then several strategies to choose start node for every ant are constructed, the threshold E increase every iteration with $1/i$ where i is the number of the current iteration:

- 1 If $E_j(i) > E$ then the subset j is forbidden for current iteration and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 2 If $E_j(i) > E$ then the subset j is forbidden for current simulation and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 3 If $E_j(i) > E$ then the subset j is forbidden for K_1 consecutive iterations and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 4 Let $r_1 \in [R, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ a node is chosen randomly from subset $\{j \mid D_j(i) > D\}$, otherwise a node is chosen randomly from the not forbidden subsets, R is chosen and fixed at the beginning.
- 5 Let $r_1 \in [R, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ a node is randomly chosen from subset $\{j \mid D_j(i) > D\}$, otherwise a node is randomly chosen from the not forbidden subsets, R is chosen at the beginning and increase with r_3 every iteration.

Where $0 \leq K_1 \leq$ "number of iterations" is a parameter. If $K_1 = 0$, than strategy 3 is equal to the random choice of the start node. If $K_1 = 1$, than strategy 3 is equal to the strategy 1. If $K_1 =$ "maximal number of iterations", than strategy 3 is equal to the strategy 2. The strategies 1, 2 and 3 can be called forbid strategies, and strategies 4 and 5 can be called stimulate strategies. By stimulate strategies the ants are forced to start there search from subsets with high value of $D_j(i)$. If $R = 0.5$, than the probability an ant to start from nodes subset with high value of $D_j(i)$ is two times high than to start from other subset.

More than one strategy for choosing the start node can be used, but there are strategies which can not be combined. The strategies are distributed into two sets: $St1 = \{strategy1, strategy2, strategy3\}$ and $St2 = \{strategy4, strategy5\}$. The strategies from same set can not be used at once. Thus it can be used strategy from one set or combine it with strategies from other set. Exemplary combinations are $(strategy1)$, $(strategy2; strategy5)$, $(strategy3; strategy4)$.

3 Experimental Results

The start strategy performance is analyzed in this section. Like test is used Multiple Knapsack Problem (MKP) as it is well-known subset problem. The Multiple Knapsack Problem has numerous applications in theory as well as in practice. It also arise as a subproblem in several algorithms for more complex problems and these algorithms will benefit from any improvement in the field

of MKP. The following major applications can be mentioned: problems in cargo loading, cutting stock, bin-packing, budget control and financial management may be formulated as MKP. In [8] is proposed to use the MKP in fault tolerance problem and in [1] is designed a public cryptography scheme whose security realize on the difficulty of solving the MKP. In [6] is mentioned that two-processor scheduling problems may be solved as a MKP. Other applications are industrial management, naval, aerospace, computational complexity theory.

The MKP can be thought as a resource allocation problem, where there are m resources (the knapsacks) and n objects and every object j has a profit p_j . Each resource has its own budget c_j (knapsack capacity) and consumption r_{ij} of resource i by object j . The aim is maximizing the sum of the profits, while working with a limited budget.

The MKP can be formulated as follows:

$$\begin{aligned} & \max \sum_{j=1}^n p_j x_j \\ & \text{subject to } \sum_{j=1}^n r_{ij} x_j \leq c_i \quad i = 1, \dots, m \\ & x_j \in \{0, 1\} \quad j = 1, \dots, n \end{aligned} \tag{7}$$

x_j is 1 if the object j is chosen and 0 otherwise.

There are m constraints in this problem, so MKP is also called m -dimensional knapsack problem. Let $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, with $c_i \geq 0$ for all $i \in I$. A well-stated MKP assumes that $p_j > 0$ and $r_{ij} \leq c_i \leq \sum_{j=1}^n r_{ij}$ for all $i \in I$ and $j \in J$. Note that the $[r_{ij}]_{m \times n}$ matrix and $[c_i]_m$ vector are both non-negative.

In the MKP one is not interested in solutions giving a particular order. Partial solution is represented by $S = \{i_1, i_2, \dots, i_j\}$ and the most recent elements incorporated to S , i_j need not be involved in the process for selecting the next element. Moreover, solutions for ordering problems have a fixed length as one search for a permutation of a known number of elements. Solutions for MKP, however, do not have a fixed length. The graph of the problem is defined as follows: the nodes correspond to the items, the arcs fully connect nodes. Fully connected graph means that after the object i one can chooses the object j for every i and j if there are enough resources and object j is not chosen yet.

The computational experience of the ACO algorithm is shown using 10 MKP instances from "OR-Library" available within WWW access at <http://people.brunel.ac.uk/~mastjjb/jeb/orlib>, with 100 objects and 10 constraints. To provide a fair comparison for the above implemented ACO algorithm, a predefined number of iterations, $k = 100$, is fixed for all the runs. The developed technique has been coded in C++ language and implemented on a Pentium 4 (2.8 Ghz). The parameters are fixed as follows: $\rho = 0.5$, $a = 1$, $b = 1$, number of used ants is 20, $A = 30$, $B = 30$, $D = 1.5$, $E = 0.5$, $K_1 = 5$, $R = 0.5$, $r_3 = 0.01$. The values of ACO parameters (ρ, a, b) are from [4] and experimentally is found that they are best for MKP. The tests are run with 1, 2, 4, 5 and 10 nodes within the nodes subsets. For every experiment, the results are obtained by

performing 30 independent runs, then averaging the fitness values obtained in order to ensure statistical confidence of the observed difference. The computational time which takes start strategies is negligible with respect to running time of the algorithm.

Tests with all combinations of strategies and with random start (12 combinations) are run. Thus the all tests are 18 000. One can observe that sometimes all nodes subsets become forbidden and the algorithm stops before performing all iterations (strategies 1, 2, 3 and combinations with them). When the nodes subsets consists of 10 nodes the algorithm does not perform all iterations for 80 of the strategies for 10 problems. When the nodes subsets consist of 5 nodes they are 36, for 4 nodes they are 30, for 2 nodes they are 21 and for 1 node they are 0. In this situation there are two possibilities. The first is to report the achieved result when the algorithm stops. The second possibility is to continue the algorithm without any strategy, applying only random start. The second possibility improves the achieved results with respect to the first one, so if all nodes subsets become forbidden the algorithm continue without any strategy.

Average achieved result by some strategy, is better than without any strategy, for every test problem. Regarding the number of the nodes in the subsets, the best average result is 1 time when they consist of 4 nodes, 6 times when they consist 2 nodes and 3 times when they consist of 1 node. The worst average result is when the algorithm is without any strategy or when the subsets consists of 10 nodes.

One can compare the achieved average results by different strategies. The achieved results by strategies 1, 2 and 3 are statistically equal, there fore latter it will be mentioned only strategy 1. For fair comparison, the difference d between the worst and best average result for every problem is divided to 10. If the average result for some strategy is between the worst average result and worst average plus $d/10$ it is appreciated with 1. If it is between the worst average plus $d/10$ and worst average plus $2d/10$ it is appreciated with 2 and so on. If it is between the best average minus $d/10$ and the best average, it is appreciated with 10. Thus for a test problem the achieved results for every strategy and every nodes deviation is appreciated from 1 to 10. After that is summed the rate of all test problems for every strategy and every nodes deviation. So theirs rate becomes between 10 and 100 (see Table 1). It is histogram like representation.

Regarding the strategies (rows) is observed that for most of them the highest rate is when the nodes subsets consists of 2 nodes. When the nodes subsets consist of 10 nodes the rate is low. The highest rate (95) have strategy combination 1-4 with two nodes in the nodes subsets and strategy combination 1-5 with 1 node in the nodes subsets. The best found average result is three times with strategy combination 1-4 with 2 nodes in the nodes subsets and never with strategy combination 1-5 with 1 node in the nodes subsets. So, the conclusion is that these two strategies/node-devision are statistically similar, but strategy combination 1-4 is slightly better.

Table 1. Estimaton of strategies and nodes deviation

number nodes	10	5	4	2	1
random	28	28	28	28	28
strat. 1	25	40	59	92	89
strat. 2	25	40	59	92	89
strat. 3	25	40	59	92	89
strat. 4	83	85	86	89	93
strat. 5	73	86	88	93	89
strat. 1-4	23	51	68	90	95
strat. 1-5	29	46	61	95	90
strat. 2-4	23	51	68	90	95
strat. 2-5	29	46	61	95	90
strat. 3-4	23	51	68	90	95
strat. 3-5	29	46	61	95	90

4 Conclusion

This paper is addressed to ant colony optimization algorithm with controlled start combining five start strategies. So, the start node of each ant depends of the goodness of the respective region. The achieved solutions with strategies are better than random start. The future work will be focused on parameter settings which manage the starting procedure. It will be investigated on influence of the parameters to algorithm performance. The aim is to study in detail the relationships between the start nodes and the quality of the achieved solutions.

Acknowledgments. This work has been partially supported by the Bulgarian National Scientific Fund under the grants ID-Modeling Processes with fixed development rules DID 02/29 and TK-Effective Monte Carlo Methods for large-scale scientific problems DTK 02/44.

References

1. Diffe, W., Hellman, M.E.: New direction in cryptography. IEEE Trans. Inf. Theory IT-36, 644–654 (1976)
2. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation 1, 53–66 (1997)
3. Fidanova, S.: Evolutionary algorithm for multiple knapsack problem. In: Int. Conference Parallel Problems Solving from Nature, Real World Optimization Using Evolutionary Computing, Granada, Spain (2002) ISB 0-9543481-0-9

4. Fidanova, S.: Ant colony optimization and multiple knapsack problem. In: Renard, J.P. (ed.) *Handbook of Research on Nature Inspired Computing for Economics and Management*, pp. 498–509. Idea Grup Inc., USA (2006)
5. Lessing, L., Dumitrescu, I., Stutzle, T.: A comparison between ACO algorithms for the set covering problem. In: *ANTS Workshop*, pp. 1–12 (2004)
6. Martello, S., Toth, P.: A mixtures of dynamic programming and branch-and-bound for the subset-sum problem. *Management Science* 30, 756–771 (1984)
7. Reiman, M., Laumanns, M.: A hybrid ACO algorithm for the capacitate minimum spanning tree problem. In: *Workshop on Hybrid Metaheuristics*, Valencia, Spain, pp. 1–10 (2004)
8. Sinha, A., Zoltner, A.A.: The multiple-choice knapsack problem. *J. Operational Research* 27, 503–515 (1979)
9. Stutzle, T., Dorigo, M.: ACO algorithm for the traveling salesman problem. In: Miettinen, K., Makela, M., Neittaanmaki, P., Periaux, J. (eds.) *Evolutionary Algorithms in Engineering and Computer Science*, pp. 163–183. Wiley, Chichester (1999)
10. Zhang, T., Wang, S., Tian, W., Zhang, Y.: ACO-VRPTWRV: A new algorithm for the vehicle routing problems with time windows and re-used vehicles based on ant colony optimization. In: *Conference on Intelligent Systems Design and Applications*, pp. 390–395. IEEE press, Los Alamitos (2006)

Sensitivity Analysis of ACO Start Strategies for Subset Problems

Stefka Fidanova¹, Pencho Marinov¹, and Krassimir Atanassov²

¹ IPP – Bulgarian Academy of Sciences, Acad. G. Bonchev str. bl.25A,
1113 Sofia, Bulgaria

{stefka,pencho}@parallel.bas.bg

² CLBME – Bulgarian Academy of Science, Acad. G. Bonchev str, bl 105,
1113 Sofia, Bulgaria

krat@bas.bg

Abstract. Ant Colony Optimization (ACO) has been used successfully to solve hard combinatorial optimization problems. This metaheuristic method is inspired by the foraging behavior of ant colonies, which manage to establish the shortest routes to feeding sources and back. On this work we use estimation of start nodes with respect to the quality of the solution. Various start strategies are offered. Sensitivity analysis of the algorithm behavior according strategy parameters is made. Our ideas is applied on Multiple Knapsack Problem (MKP) like a representative of the subset problems.

1 Introduction

Many combinatorial optimization problems are fundamentally hard. This is the most typical scenario when it comes to realistic and relevant problems in industry and science. Linear programming and dynamic programming techniques, for example, often fail in solving NP-hard problems with large number of variables. Examples of optimization problems are Traveling Salesman Problem [11], Vehicle Routing [13], Minimum Spanning Tree [9], Multiple Knapsack Problem [5], etc. They are NP-hard problems and in order to obtain solution close to the optimality in reasonable time, metaheuristic methods are used. One of them is Ant Colony Optimization (ACO) [3].

Real ants foraging for food lay down quantities of pheromone (chemical cues) marking the path that they follow. An isolated ant moves essentially at random but an ant encountering a previously laid pheromone will detect it and decide to follow it with high probability and thereby reinforce it with a further quantity of pheromone. The repetition of the above mechanism represents the auto-catalytic behavior of a real ant colony where the more the ants follow a trail, the more attractive that trail becomes.

ACO is inspired by real ant behavior to solve hard combinatorial optimization problems. The ACO algorithm uses a colony of artificial ants that behave as cooperative agents in a mathematical space where they are allowed to search and reinforce pathways (solutions) in order to find the optimal ones. The problem is represented by graph and the ants walk on the graph to construct solutions.

The solutions are represented by paths in the graph. After the initialization of the pheromone trails, the ants construct feasible solutions, starting from random nodes, and then the pheromone trails are updated. At each step the ants compute a set of feasible moves and select the best one (according to some probabilistic rules) to continue the rest of the tour.

The structure of the ACO algorithm is shown by the pseudocode below (Figure 1).

```

Ant Colony Optimization
Initialize number of ants;
Initialize the ACO parameters;
while not end-condition do
    for k=0 to number of ants
        ant k chose start node;
        while solution is not constructed do
            ant k selects higher probability node;
        end while
    end for
    Update-pheromone-trails;
end while

```

Fig. 1. Pseudocode for ACO

The transition probability $p_{i,j}$, to choose the node j when the current node is i , is based on the heuristic information $\eta_{i,j}$ and the pheromone trail level $\tau_{i,j}$ of the move, where $i, j = 1, \dots, n$.

$$p_{i,j} = \frac{\tau_{i,j}^a \eta_{i,j}^b}{\sum_{k \in Unused} \tau_{i,k}^a \eta_{i,k}^b}, \quad (1)$$

where *Unused* is the set of unused nodes of the graph.

The higher the value of the pheromone and the heuristic information, the more profitable it is to select this move and resume the search. In the beginning, the initial pheromone level is set to a small positive constant value τ_0 ; later, the ants update this value after completing the construction stage. ACO algorithms adopt different criteria to update the pheromone level. The pheromone trail update rule is given by:

$$\tau_{i,j} \leftarrow \rho \tau_{i,j} + \Delta \tau_{i,j}, \quad (2)$$

where ρ models evaporation in the nature and $\Delta \tau_{i,j}$ is new added pheromone which is proportional to the quality of the solution.

The novelty in this work is the use of estimation of start nodes with respect to the quality of the solution and thus to better manage the search process. Various start strategies and their combinations are offered. Sensitivity analysis of the algorithm according strategy parameters is made. Our ideas is applied on Multiple Knapsack Problem like a representative of the subset problems.

The rest of the paper is organized as follows: in section 2 several start strategies are proposed. In section 3 the MKP is introduced. In section 4 the strategies are applied on MKP and sensitivity analysis of the algorithm according strategy parameters is made. At the end some conclusions and directions for future work are done.

2 Start Strategies

The known ACO algorithms create a solution starting from random node. But for some problems, especially subset problems, it is important from which node the search process starts. For example if an ant starts from node which does not belong to the optimal solution, probability to construct it is zero. In this paper is offered several start strategies. The aim is to use the experience of the ants from previous iteration to choose the better starting node. Other authors use this experience only by the pheromone, when the ants construct the solutions. Therefore several start strategies are offered.

Let the graph of the problem has m nodes. The set of nodes is divided on N subsets. There are different ways for dividing. Normally, the nodes of the graph are randomly enumerated. An example for creating of the subsets, without lost of generality, is: the node number one is in the first subset, the node number two - in the second subset, etc., the node number N is in the N -th subset, the node number $N + 1$ is in the first subset, etc. Thus the number of nodes in the separate subsets are almost equal. After the first iteration the estimations $D_j(i)$ and $E_j(i)$ are introduced of the node subsets, where $i \geq 2$ is the number of the current iteration and $D_j(i)$ and $E_j(i)$ are weight coefficients of j -th node subset ($1 \leq j \leq N$), which are calculated by the following formulas:

$$D_j(i) = \frac{i \cdot D_j(i-1) + F_j(i)}{i}, \quad (3)$$

$$E_j(i) = \frac{i \cdot E_j(i-1) + G_j(i)}{i}, \quad (4)$$

where $i \geq 2$ is the current iteration and for each j ($1 \leq j \leq N$):

$$F_j(i) = \begin{cases} \frac{f_{j,A}}{n_j} & \text{if } n_j \neq 0 \\ F_j(i-1) & \text{otherwise} \end{cases}, \quad (5)$$

$$G_j(i) = \begin{cases} \frac{g_{j,B}}{n_j} & \text{if } n_j \neq 0 \\ G_j(i-1) & \text{otherwise} \end{cases}, \quad (6)$$

and $f_{j,A}$ is the number of the solutions among the best $A\%$, and $g_{j,B}$ is the number of the solutions among the worst $B\%$, where $A + B \leq 100$, $i \geq 2$ and

$\sum_{j=1}^N n_j = n$, where n_j ($1 \leq j \leq N$) is the number of solutions obtained by ants starting from nodes subset j . Initial values of the weight coefficients are : $D_j(1) = 1$ and $E_j(1) = 0$. Obviously $F_j(i)$ and $G_j(i) \in [0, 1]$.

Let threshold E for $E_j(i)$ and D for $D_j(i)$ be fixed, than several strategies to choose start node for every ant are constructed, the threshold E increase every iteration with $1/i$ where i is the number of the currently iteration:

- 1 If $\frac{E_j(i)}{D_j(i)} > E$ then the subset j is forbidden for **current iteration** and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 2 If $\frac{E_j(i)}{D_j(i)} > E$ then the subset j is forbidden for **current simulation** and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 3 If $\frac{E_j(i)}{D_j(i)} > E$ then the subset j is forbidden for K_1 **consecutive iterations** and the starting node is chosen randomly from $\{j \mid j \text{ is not forbidden}\}$;
- 4 Let $r_1 \in [R, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ a node is chosen randomly from subset $\{j \mid D_j(i) > D\}$, otherwise a node is chosen randomly from the not forbidden subsets, R is chosen and fixed at the beginning.
- 5 Let $r_1 \in [R, 1)$ is a random number. Let $r_2 \in [0, 1]$ is a random number. If $r_2 > r_1$ a node is randomly chosen from subset $\{j \mid D_j(i) > D\}$, otherwise a node is randomly chosen from the not forbidden subsets, R is chosen at the beginning and increases with r_3 every iteration.

Where $0 \leq K_1 \leq$ "number of iterations" is a parameter. If $K_1 = 0$, than strategy 3 is equal to the random choose of the start node. If $K_1 = 1$, than strategy 3 is equal to the strategy 1. If $K_1 =$ "maximal number of iterations", than strategy 3 is equal to the strategy 2. The strategies 1, 2 and 3 can be called forbid strategies, and strategies 4 and 5 can be called stimulate strategies. By stimulate strategies the ants are forced to start there search from subsets with high value of $D_j(i)$. If $R = 0.5$, than the probability an ant to start from nodes subset with high value of $D_j(i)$ is two times high than to start from other subset. For forbidden strategies is used fraction between $E_j(i)$ and $D_j(i)$. Thus is prevented some regions with several bad and with several good solutions to be forbidden.

More than one strategy for choosing the start node can be used, but there are strategies which can not be combined. The strategies are distributed into two sets: $St1 = \{strategy1, strategy2, strategy3\}$ and $St2 = \{strategy4, strategy5\}$. The strategies from same set can not be used at once. Thus it can be used strategy from one set or combine it with strategies from other set. Exemplary combinations are $(strategy1)$, $(strategy2; strategy5)$, $(strategy3; strategy4)$.

3 Multiple Knapsack Problem

We test the ideas for controlled start on MKP. MKP is a real world problem and is a representative of the class of subset problems. The MKP has numerous applications in theory as well as in practice. It also arise as a subproblem in several algorithms for more complex problems and these algorithms will benefit from

any improvement in the field of MKP. The following major applications can be mentioned: problems in cargo loading, cutting stock, bin-packing, budget control and financial management. Sinha and Zoltner [10] proposed to use the MKP in fault tolerance problem and in [2] is designed a public cryptography scheme whose security realize on the difficulty of solving the MKP. Martello and Toth [8] mention that two-processor scheduling problems may be solved as a MKP. Other applications are industrial management, naval, aerospace, computational complexity theory.

The MKP can be thought as a resource allocation problem, where there are m resources (the knapsacks) and n objects and every object j has a profit p_j . Each resource has its own budget c_j (knapsack capacity) and consumption r_{ij} of resource i by object j . The aim is maximizing the sum of the profits, while working with a limited budget.

The MKP can be formulated as follows:

$$\begin{aligned} & \max \sum_{j=1}^n p_j x_j \\ & \text{subject to } \sum_{j=1}^n r_{ij} x_j \leq c_i \quad i = 1, \dots, m \\ & x_j \in \{0, 1\} \quad j = 1, \dots, n \end{aligned} \tag{7}$$

x_j is 1 if the object j is chosen and 0 otherwise.

There are m constraints in this problem, so MKP is also called m -dimensional knapsack problem. Let $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, with $c_i \geq 0$ for all $i \in I$. A well-stated MKP assumes that $p_j > 0$ and $r_{ij} \leq c_i \leq \sum_{j=1}^n r_{ij}$ for all $i \in I$ and $j \in J$. Note that the $[r_{ij}]_{m \times n}$ matrix and $[c_i]_m$ vector are both non-negative.

In the MKP one is not interested in solutions giving a particular order. Therefore a partial solution is represented by $S = \{i_1, i_2, \dots, i_j\}$ and the most recent elements incorporated to S , i_j need not be involved in the process for selecting the next element. Moreover, solutions for ordering problems have a fixed length as one search for a permutation of a known number of elements. Solutions for MKP, however, do not have a fixed length. The graph of the problem is defined as follows: the nodes correspond to the items, the arcs fully connect nodes. Fully connected graph means that after the object i one can chooses the object j for every i and j if there are enough resources and object j is not chosen yet.

4 Experimental Results

Sensitivity analysis of the algorithm according strategy parameter K_1 is made in this section. The computational experience of the ACO algorithm is shown using 10 MKP instances from “OR-Library” available within WWW access at <http://people.brunel.ac.uk/mastjjb/jeb/orlib/>, with 100 objects and 10 constraints. To provide a fair comparison for the above implemented ACO algorithm, a predefined number of iterations, $k = 100$, is fixed for all the runs.

If the value of k (number of iterations) is too high, the achieved results will be very close to the optimal solution and will be difficult to appreciate different strategies. We apply strategies on MMAS [12], because it is one of the best ACO approach. The developed technique has been coded in C++ language and implemented on a Pentium 4 (2.8 Ghz). The parameters are fixed as follows: $\rho = 0.5$, $a = 1$, $b = 1$, number of used ants is 20, $A = 30$, $B = 30$, $D = 1.5$, $E = 0.5$, $R = 0.5$, $r_3 = 0.01$. The values of ACO parameters (ρ, a, b) are from [6] and experimentally is found that they are best for MKP. The tests are run with 1, 2, 4, 5 and 10 nodes within the nodes subsets and values for K_1 are 1, 2, 5, 10, 20, 25, 50 and 100. For every experiment, the results are obtained by performing 30 independent runs, then averaging the fitness values. The computational time which takes start strategies is negligible with respect to the computational time which takes solution construction.

Tests with strategies 3, 3-4, 3-5 and with random start (they are strategies concerning parameter K_1), eight values for K_1 and five kind of node subsets are run and every test is run 30 times for comparing reason. Thus the all runs are 72 030. One can observe that sometimes all nodes subsets become forbidden and the algorithm stops before performing all iterations. So if all nodes subsets become forbidden the algorithm performs several iterations without any strategy with random start till some of the subsets become not forbidden. Then the algorithm continue to apply the chosen strategy.

The problem which arise is how to compare the achieved solutions by different strategies and different node-divisions. Therefore the difference (interval) d between the worst and best average result for every problem is divided to 10. If the average result for some strategy, node division and K_1 is in the first interval with borders the worst average result and worst average plus $d/10$ it is appreciated with 1. If it is in the second interval with borders the worst average plus $d/10$ and worst average plus $2d/10$ it is appreciated with 2 and so on. If it is in the 10th interval with borders the best average minus $d/10$ and the best average result, it is appreciated with 10. Thus for a test problem the achieved results for every strategy, every nodes division and every K_1 is appreciated from 1 to 10. After that is summed the rate of all test problems for every strategy, every nodes division and K_1 . So the rate of the strategies/node-division/ K_1 becomes between 10 and 100, because the benchmark problems are 10. It is mode of result classification. The best results are achieved when the node subsets consists of two nodes, therefore we will report only them [7].

Analysing Table 1 we observe that poorest results are without any strategy. Regarding influence of the parameter K_1 the rate is higher when $K_1 \leq 20$. We can conclude that it is better the node subsets to be forbidden for small number of iterations. If the parameter K_1 has a big value and some node subset is forbidden at the beginning of the simulation, than may be good solutions start from this subset too, but it will not be investigated and will not be appreciated properly. Therefore we recommend the value of parameter K_1 to be small, up to 20.

Table 1. Estimation of strategies and nodes division

K	1	2	5	10	20	25	50	100
average	10	10	10	10	10	10	10	10
strat 3	82	80	83	88	88	83	85	85
start 3-4	90	89	90	87	89	88	88	88
start 3-5	88	89	89	87	88	86	86	86

5 Conclusion

On this paper we apply start strategies on ACO algorithm for MKP. We make sensitivity analysis of the strategy parameter K , number of iterations on which the node subsets stay forbidden. We test our ideas on 10 test problems. After analysing the result our conclusion is that the parameter value need to be small. For future work we will apply other start strategies and we will make sensitivity analysis of the algorithm according strategy parameters.

Acknowledgements. This work has been partially supported by the Bulgarian National Scientific Fund under the grants ID-Modelling Processes with fixed development rules DID 02/29 and TK-Effective Monte Carlo Methods for large-scale scientific problems DTK 02/44.

References

1. Bonabeau, E., Dorigo, M., Theraulaz, G.: *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York (1999)
2. Diffe, W., Hellman, M.E.: New direction in cryptography. *IEEE Trans Inf. Theory* IT-36, 644–654 (1976)
3. Dorigo, M., Gambardella, L.M.: Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Transactions on Evolutionary Computation* 1, 53–66 (1997)
4. Dorigo, M., Stutzle, T.: *Ant Colony Optimization*. MIT Press, Cambridge (2004)
5. Fidanova, S.: Evolutionary Algorithm for Multiple Knapsack Problem. In: *Int. Conference Parallel Problems Solving from Nature, Real World Optimization Using Evolutionary Computing*, Granada, Spain (2002) ISBN 0-9543481-0-9
6. Fidanova, S.: Ant colony optimization and multiple knapsack problem. In: Renard, J.P. (ed.) *Handbook of Research on Nature Inspired Computing for Economics and Management*, pp. 498–509. Idea Grup Inc., USA (2006) ISBN 1-59140-984-5
7. Fidanova, S., Atanassov, K., Marinov, P., Parvathi, R.: Ant Colony Optimization for Multiple Knapsack Problems with Controlled Starts. *Int. J. Bioautomation* 13(4), 271–280
8. Martello, S., Toth, P.: A mixtures of dynamic programming and branch-and-bound for the subset-sum problem. *Management Science* 30, 756–771 (1984)
9. Reiman, M., Laumanns, M.: A Hybrid ACO algorithm for the Capacitated Minimum Spanning Tree Problem. In: *Proc. of First Int. Workshop on Hybrid Metaheuristics*, Valencia, Spain, pp. 1–10 (2004)

10. Sinha, A., Zoltner, A.A.: The multiple-choice knapsack problem. *J. Operational Research* 27, 503–515 (1979)
11. Stutzle, T., Dorigo, M.: ACO Algorithm for the Traveling Salesman Problem. In: Miettinen, K., Makela, M., Neittaanmaki, P., Periaux, J. (eds.) *Evolutionary Algorithms in Engineering and Computer Science*, pp. 163–183. Wiley, Chichester (1999)
12. Stutzle, T., Hoos, H.H.: MAX-MIN Ant System. In: Dorigo, M., Stutzle, T., Di Caro, G. (eds.) *Future Generation Computer Systems*, vol. 16, pp. 889–914 (2000)
13. Zhang, T., Wang, S., Tian, W., Zhang, Y.: ACO-VRPTWRV: A New Algorithm for the Vehicle Routing Problems with Time Windows and Re-used Vehicles based on Ant Colony Optimization. In: *Sixth International Conference on Intelligent Systems Design and Applications*, pp. 390–395. IEEE press, Los Alamitos (2006)

A Highly-Parallel TSP Solver for a GPU Computing Platform

Noriyuki Fujimoto¹ and Shigeyoshi Tsutsui²

¹ Osaka Prefecture University, 1-1 Gakuen-Cho, Naka-ku, Sakai-shi,
Osaka, 599-8531, Japan

`fujimoto@mi.s.osakafu-u.ac.jp`

² Hannan University, 5-4-33 Amamihigashi, Matsubara,
Osaka, 580-8502, Japan

`tsutsui@hannan-u.ac.jp`

Abstract. The traveling salesman problem (TSP) is probably the most widely studied combinatorial optimization problem and has become a standard testbed for new algorithmic ideas. Recently the use of a GPU (Graphics Processing Unit) to accelerate non-graphics computations has attracted much attention due to its high performance and low cost. This paper presents a novel method to solve TSP with a GPU based on the CUDA architecture. The proposed method highly parallelizes a serial metaheuristic algorithm which is a genetic algorithm with the OX (order crossover) operator and the 2-opt local search. The experiments with an NVIDIA GeForce GTX285 GPU and a single core of 3.0 GHz Intel Core2 Duo E6850 CPU show that our GPU implementation is about up to 24.2 times faster than the corresponding CPU implementation.

Keywords: parallel metaheuristic, genetic algorithm, GPGPU.

1 Introduction

The traveling salesman problem [1,8] (TSP) is probably the most widely studied combinatorial optimization problem and has become a standard testbed for new algorithmic ideas [6]. Recently the use of a GPU (Graphics Processing Unit) to accelerate non-graphics computations has attracted much attention due to its high performance and low cost.

This paper presents a novel method to solve TSP with a GPU based on the CUDA architecture [10]. Especially for CUDA, the proposed method highly parallelizes a serial metaheuristic algorithm which is a genetic algorithm (GA for short) with the OX (order crossover) operator [5] and the 2-opt local search [4].

Genetic algorithms have obvious parallelism among individuals. However, the parallelism is not enough to obtain high performance of a GPU. To utilize an advantage of "many-thread" architecture of CUDA, we extract not only the parallelism among individuals but also another parallelism in the processing of each individual. That is, we parallelize the execution of each OX operator and each 2-opt local search, too.

To evaluate the effectiveness of the proposed method, we conduct some experiments for TSPLIB [13] benchmark problem instances using an NVIDIA GeForce GTX285 GPU and a 3.0 GHz Intel Core 2 Duo E6850 CPU. The experimental results show that our GPU implementation is about up to 24.2 times faster than the corresponding CPU implementation.

Quite recently, CUDA has been successfully used to accelerate various applications in scientific fields such as fluid dynamics, image processing, and simulations. However, in the field of genetic algorithms to solve TSP, almost no result is known except for Sanci's result [14]. This is because genetic algorithms have special property of frequent random access to large data-structures, which is not the case in the other successful fields. Sanci achieved at most 4.9 times speedup using an NVIDIA GeForce 9800M GPU and a 2.13GHz Intel Core2 Duo P8400 CPU.

The remainder of this paper is organized as follows. Section 2 presents the proposed algorithm. Experiments to show the performance of the proposed algorithm are reported in Section 3. Section 4 gives some concluding remarks and future works. Due to the limited space, this paper include no description on CUDA. Readers unfamiliar with CUDA are recommended the literature [10,12].

2 The Proposed Algorithm

2.1 An Overview of the Proposed Algorithm

Listing 1.1 shows a pseudo code of a serial GA program to solve a TSP instance with the OX operator and 2-opt local search. Usually, a crossover operator generates two offspring from two parents. However, in the proposed method, we generate only one child from two parents. In Step 18 of Listing 1.1, the comparison of costs is performed like a tournament selection with size 2. However, each comparison is performed between individuals $s1[i]$ and $s2[i]$ which have the same index i . Please remember here that $s2[i]$ is generated from $s1[i]$ as one of its parents (the other parent $s1[j]$ was chosen randomly). Since a parent and a child have partly similar substrings, this comparison scheme can be expected to maintain population diversity like the deterministic crowding proposed by Mahfoud [11]. Using one child from two parents was already proposed for designing the well known GENITOR algorithm by Whitley et al. [16].

Genetic algorithms have obvious parallelism among individuals. So, also in the case of the serial GA program, **for** loops in line 5, 11, and 16 can be executed in parallel. Before parallelizing the serial GA program, we conducted some preliminary experiments to investigate how many individuals are required to solve a TSP instance efficiently. The results show that at most tens of individuals are enough for instances at most 500 cities from TSPLIB benchmark. On the CUDA architecture, the only way to hide memory access latency is to execute other threads when some threads are stalled due to memory access latency [12]. At a time, a CUDA GPU can make 30720 threads of all the running threads active. So, to hide memory access latency efficiently, at least hundred thousands of threads should be created. Therefore, the parallelism among individuals is not sufficient at all to extract high performance of a GPU.

Listing 1.1. A serial GA with OX operator and 2-opt local search for TSP

```

1 // s1[0 .. n - 1] and p1[0 .. n - 1] :
2 // p1[i] is the length of tour s1[i]
3 // s2[0 .. n - 1] : buffer for individuals generated by
4 // OX and 2-opt
5 // p2[0 .. n - 1] : p2[i] is the length of tour s2[i]
6 generation = 0;
7 for (i = 0; i < populationSize; i++) {
8     s1[i] = a random tour;
9     p1[i] = evaluate(s1[i]);
10    if (p1[i] is acceptable) return i; // found solution is s1[i]
11 }
12 while (generation < maxGeneration) {
13     for (i = 0; i < populationSize; i++) {
14         j = a random integer such that
15         0 <= j < populationSize and j != i;
16         s2[i] = OX(s1[i], s1[j]);
17         2OPT_best(s2[i]);
18     }
19     for (i = 0; i < populationSize; i++) {
20         p2[i] = evaluate(s2[i]);
21         if (p2[i] < p1[i]) {
22             s1[i] = s2[i];
23             p1[i] = p2[i];
24         }
25         if (p1[i] is acceptable) return i; // found solution is s1[i]
26     }
27     generation++;
28 }
29 return -1; // solution is not found

```

Hence, the proposed method parallelizes the serial GA program furthermore. That is, we also parallelized the OX operator and 2-opt local search. To realize this approach, the proposed method adopted the following structure:

- The proposed method runs m thread blocks concurrently where m is the number of individuals.
- Each individual is processed by a thread block with n threads where n is the number of cities of a given TSP instance.
- Each thread block concurrently performs the OX operator and 2-opt local search in the best improvement manner [6] for its individual.

Listing 1.2 shows a high level description of the proposed method.

The 2-opt local search in the best improvement manner can be easily parallelized. So, the next subsection describes how the proposed method parallelizes the OX operator.

2.2 Parallelization of the OX Operator

OX constructs an offspring by choosing a subsequence of one parent and preserving the relative order of cities of the other parent. Listing 1.3 shows an implementation of OX in the C programming language. The time complexity of OX is $O(n^2)$ where n is the number of cities. It is not so clear whether OX has

Listing 1.2. A high level description of the proposed method

```

1  --global-- void kernel(int *s, int *d, int *ps, int *pd,
2                      int *found)
3  {
4      int i = blockIdx.x;
5
6      if (threadIdx.x == 0) {
7          j = a random integer such that
8              0 <= j < populationSize and j != i;
9      }
10     __syncthreads();
11
12     // s_si, s_sj, s_pi, s_pj, and s_d are pointers to shared mem.
13     copy s[i] to *s_si in parallel within this block;
14     copy s[j] to *s_sj in parallel within this block;
15     if (threadIdx.x == 0) copy ps[i] to *s_pi;
16     *s_d = OX(s_si, s_sj) in parallel within this block;
17     2OPT_best(*s_d) in parallel within this block;
18     p = evaluate(s_d) in parallel within this block;
19
20     __syncthreads();
21     if (p < *s_pi) {
22         d[i] = *s_d in parallel within this block;
23         if (threadIdx.x == 0) pd[i] = p;
24     } else {
25         d[i] = *s_si in parallel within this block;
26         if (threadIdx.x == 0) p = pd[i] = ps[i];
27     }
28     if (threadIdx.x == 0) {
29         if (p is acceptable) *found = i; // found solution is d[i]
30     }
31 }
32
33 int main()
34 {
35     // s1, s2, p1, and p2 are pointers to main memory.
36     // d_s1, d_s2, d_p1, d_p2, and d_found are pointers to VRAM.
37     generation = 0;
38     for (i = 0; i < populationSize; i++) {
39         s1[i] = a random tour; p1[i] = evaluate(s1[i]);
40         if (p1[i] is acceptable) return i; // found solution is s1[i]
41     }
42     copy s1[*] to d_s1[*]; copy p1[*] to d_p1[*];
43     while (generation < maxGeneration) {
44         set *d_found to -1;
45         int grid = populationSize; int block = the number of cities;
46         kernel<<< grid, block >>>(d_s1, d_s2, d_p1, d_p2, d_found);
47         if (*d_found >= 0) { acceptable solution is found; break; }
48         swap(d_s1, d_s2); // swap pointers only (double buffering)
49         swap(d_p1, d_p2); // swap pointers only (double buffering)
50         generation++;
51     }
52 }

```

parallelism. So, we propose a parallelized OX. Our parallelized OX generates the same computational results as the original OX. Listing 1.4 shows our parallelization of OX in the C programming language with 'parallel for' construct to indicate parallelism. The total computational amount of our parallelizable OX is the same as the serial time complexity of ordinary OX. Our parallelization uses a prefix sums operation (or scan operation) [2,3] which is a well-known

Listing 1.3. The OX operator

```

1 void OX(int n, int *s1, int *s2, int *d)
2 // s1[0 .. n - 1] and s2[0 .. n - 1] : given two individuals
3 // d[0 .. n - 1] : buffer for a generated individual
4 {
5     Set cut1 and cut2 randomly s. t.
6     0 <= cut1, cut2 < n and cut1 != cut2;
7     for(int j = 0; j < n; j++) d[j] = s1[j];
8     for(int j = cut1; j < cut2; j++){
9         for(int pos = 0; pos < n; pos++) if (s2[j] == d[pos]) break;
10        int next = pos + 1;
11        while(next != cut2){
12            if (pos == n) pos = 0;
13            if (next == n) next = 0;
14            d[pos++] = d[next++];
15        }
16    }
17    for(int j = cut1; j < cut2; j++) d[j] = s2[j];
18 }

```

Listing 1.4. OX in a parallelizable fashion

```

1 void PrefixSums(int n, int *s, int *d, int begin = 0)
2 {
3     d[begin] = s[begin];
4     for (int i = begin + 1; i < n; i++) d[i] = d[i - 1] + s[i];
5     if (begin > 0) {
6         d[0] = d[n - 1] + s[0];
7         for (int i = 1; i < begin; i++) d[i] = d[i - 1] + s[i];
8     }
9 }
10
11 void ParallelizableOX(int n, int *s1, int *s2, int *d)
12 // s1[0 .. n - 1] and s2[0 .. n - 1] : given two individuals
13 // d[0 .. n - 1] : buffer for a generated individual
14 {
15     Set cut1 and cut2 randomly s. t.
16     0 <= cut1, cut2 < n and cut1 != cut2;
17     for (int j = 0; j < n; j++) do in parallel {
18         for (int pos = cut1; pos < cut2; pos++)
19             if (s1[j] == s2[pos]) break;
20         ps[j] = (pos >= cut2) ? 1 : 0;
21     }
22     PrefixSums(n, ps, to, (cut2 < n) ? cut2 : 0);
23     for (int j = 0; j < n; j++) do in parallel {
24         to[j]--;
25         if (to[j] < n - cut2) to[j] += cut2;
26         else to[j] -= (n - cut2);
27     }
28     for (int j = 0; j < n; j++) do in parallel
29         if (ps[j]) d[to[j]] = s1[j];
30     for (int j = cut1; j < cut2; j++) do in parallel d[j] = s2[j];
31 }

```


Table 1. Comparison between the proposed CUDA program and the corresponding CPU program in case that population size is 60, the maximum number of generations is 1000, and acceptable error ratio to the optimal solution is 0.5%

problem instance	CPU			GPU			speedup ratio
	#success in 10 trials	avg. exec. time(sec)	avg. err. (%)	#success in 10 trials	avg. exec. time(sec)	avg. err. (%)	
gr120	10	5.7578	0.44	10	0.5156	0.41	11.2
pr124	10	3.3625	0.23	10	0.2297	0.30	14.6
bier127	10	5.6109	0.37	10	0.4563	0.35	12.3
ch130	10	6.8593	0.28	10	0.5110	0.35	13.4
pr136	10	6.5781	0.44	10	0.4860	0.28	13.5
gr137	10	5.9515	0.27	10	0.4813	0.41	12.4
pr144	10	5.1062	0.21	10	0.2954	0.24	17.3
ch150	10	11.5516	0.37	10	0.7594	0.37	15.2
pr152	10	6.9219	0.40	10	0.4579	0.41	15.1
u159	10	9.8500	0.26	10	0.5656	0.22	17.4
brg180	10	26.7625	0.00	10	1.7281	0.00	15.5
d198	10	21.7875	0.39	10	1.4109	0.43	15.4
kroA200	10	29.6437	0.38	10	1.4891	0.40	19.9
kroB200	10	31.4468	0.38	10	1.5828	0.38	19.9
gr202	10	32.0156	0.43	10	1.9438	0.40	16.5
ts225	10	30.4671	0.38	10	1.2578	0.38	24.2
tsp225	10	45.8062	0.33	10	2.5688	0.41	17.8
pr226	10	27.1046	0.35	10	1.3860	0.39	19.6
gr229	10	49.4391	0.41	10	3.0172	0.42	16.4
gil262	10	60.4265	0.38	10	3.9079	0.41	15.5
pr264	10	34.5985	0.35	10	2.6063	0.40	13.3
a280	10	54.8969	0.35	10	4.0703	0.39	13.5
pr299	10	65.8516	0.45	10	4.7344	0.44	13.9
lin318	10	91.0219	0.41	10	6.8016	0.43	13.4
rd400	10	304.5594	0.46	10	22.8266	0.45	13.3
fl417	10	150.0360	0.47	10	11.6625	0.48	12.9
gr431	9	332.1944	0.45	9	29.9359	0.48	11.1
pr439	10	201.9344	0.45	10	17.2250	0.45	11.7
pcb442	9	330.1320	0.47	8	35.4219	0.46	9.3
d493	10	461.7297	0.46	9	48.5052	0.48	9.5

data-parallel primitive with wide variety of applications [2] and can be efficiently executed in parallel [7]. Also for CUDA GPUs, an efficient parallel implementation of prefix sums is provided by Sengupta et al. [15]. As for prefix sums operations, our implementation uses Sengupta et al.'s CUDA implementation.

3 Experiments

This section compares the performance of the proposed CUDA program with a CPU program that performs the same computation. For each test, a single core of 3.0 GHz Intel Core2 Duo E6850 and NVIDIA GeForce GTX285 was used. The

OS used is Windows XP Professional SP3 with NVIDIA graphics driver Version 195.62. For compilation, Microsoft Visual Studio 2008 Professional Edition with optimization option /O2 and CUDA 2.3 SDK were used.

Table 1 shows the performance of the proposed GPU algorithm and the corresponding CPU algorithm for problem instances from TSPLIB benchmark with at least 120 cities and at most 512 cities. TSPLIB provides the optimal lengths of tours for these problem instances. We measured the execution time for our GPU (CPU) program to find an acceptable solution. An acceptable solution is a solution with tour length within a factor of $(1 + \text{a given acceptable error ratio})$ to the optimal tour length. In our experiments, acceptable error ratio is 0.5%. For each problem instance, the measurement was conducted 10 times consecutively and the average value of successful trials of the 10 trials was adopted where successful trial is a trial such that an acceptable solution is found within 1000 generations of our GA. The speedup ratio is the ratio of the execution time of CPU to that of GPU. The speedup ratio indicates how the GPU program is faster than the CPU program.

We can see the proposed algorithm achieves 24.2 times speedup at the maximum compared with a single core of the CPU. We can also see that the performance of the CUDA program is maximized in case that the number of cities is from 200 to 226. This is because the number of active threads is roughly maximized for such number of cities. This can be verified by the fact that the implementation of the proposed method uses 25 registers per thread and $(20n + 28)$ bytes shared memory per thread block where n is the number of cities.

4 Conclusion and Future Work

A new parallel implementation has been proposed for solving the traveling salesman problem on the NVIDIA CUDA GPU architecture. The proposed algorithm is a highly parallel variant of a genetic algorithm with the OX operator and 2-opt local search. The proposed GPU algorithm has achieved 24.2 times speed up to the corresponding CPU algorithm.

Our experiments was conducted for problem instances at most 512 cities. This is mainly because 2-opt heuristic is insufficient for larger problem instances. Therefore, one of future works is to develop a CUDA implementation for much larger problem instances by parallelizing more sophisticated heuristics (e.g., Lin-Kernighan heuristic [9]) for CUDA.

References

1. Applegate, D.L., Bixby, R.E., Chvátal, V., Cook, W.J.: The Traveling Salesman Problem: A Computational Study. Princeton University Press, Princeton (2007)
2. Brelloch, G.E.: Scans as Primitive Parallel Operations. *IEEE Transactions on Computing* 38(11), 1526–1538 (1989)
3. Brelloch, G.E.: Vector Models for Data-Parallel Computing. MIT Press, Cambridge (1990)

4. Croes, G.A.: A Method for Solving Traveling Salesman Problems. *Operations Research* 6, 791–812 (1958)
5. Davis, L.: Applying Adaptive Algorithms to Epistatic Domains. In: *Proc. of the International Joint Conference on Artificial Intelligence*, pp. 162–164 (1985)
6. Hoos, H.H., Stützle, T.: *Stochastic Local Search: Foundations and Applications*. Elsevier, Amsterdam (2005)
7. JaJa, J.: *An Introduction to Parallel Algorithms*. Addison-Wesley Professional, Reading (1992)
8. Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B.: *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley, Chichester (1985)
9. Lin, S., Kernighan, B.W.: An Effective Heuristic Algorithm for the Traveling-Salesman Problem. *Operations Research* 21, 498–516 (1973)
10. Lindholm, E., Nickolls, J., Oberman, S., Montrym, J.: NVIDIA Tesla: A Unified Graphics and Computing Architecture. *IEEE Micro* 28(2), 39–55 (2008)
11. Mahfoud, S.: A Comparison of Parallel and Sequential Niching Methods. In: *Proc. of the International Conference on Genetic Algorithms*, pp. 136–143 (1995)
12. NVIDIA: *CUDA Programming Guide 3.1* (2010),
http://www.nvidia.com/object/cuda_develop.html
13. Reinelt, G.: TSPLIB: A Traveling Salesman Problem Library. *ORSA Journal on Computing* 3, 376–384 (1991)
14. Sanci, S.: *A Parallel Algorithm for Flight Route Planning on GPU Using CUDA*, Master Thesis, Middle East Technical University (April 2010)
15. Sengupta, S., Harris, M., Garland, M.: Efficient Parallel Scan Algorithms for GPUs. NVIDIA Technical Report NVR-2008-003 (2008)
16. Whitley, L.D., Starkweather, T., Fuquay, D.: Scheduling Problems and Traveling Salesman Problem: The Genetic Edge Recombination Operator. In: *Proc. of the International Conference on Genetic Algorithms*, pp. 133–140 (1989)

Metaheuristics for the Asymmetric Hamiltonian Path Problem

João Pedro Pedroso

INESC - Porto

and

DCC - Faculdade de Ciências, Universidade do Porto, Portugal

jpp@fc.up.pt

Abstract. One of the most important applications of the Asymmetric Hamiltonian Path Problem is in scheduling. In this paper we describe a variant of this problem, and develop both a mathematical programming formulation and simple metaheuristics for solving it. The formulation is based on a transformation of the input data, in such a way that a standard mathematical programming model for the Asymmetric Travelling Salesman Problem can be used on this slightly different problem. Two standard metaheuristics for the asymmetric travelling salesman are proposed and analysed on this variant: repeated random construction followed by local search with the 3-Exchange neighbourhood, and iterated local search based on the same neighbourhood and on a 4-Exchange perturbation. The computational results obtained show the interest and the complementary merits of using a mixed-integer programming solver and an approximative method for the solution of this problem.

1 Introduction

We are dealing with the following problem: given an operation currently being done in a machine, determine the order for the set of operations to be produced next, such that the total production time is minimized. There are no precedence constraints among the operations, but there are changeover times which depend on the production sequence. Minimizing the total production time is equivalent to minimizing the time spent in changeovers, as the other times are constant.

This problem is relevant in many practical situations. In paint production the machine cleaning times are usually dependent on the sequence; for example, producing white colour after grey requires a much more careful cleaning than the other way around. The production of steel is also a situation where the sequence of production is very important, having very strict rules and costs that depend on the order. Yet another practical application is in food manufacturing, where strong flavours can be produced after flavourless products at a small cost, but very careful and lengthy cleaning is required in the inverse situation.

One possibility for modelling this problem is to consider a graph with a node for each of the items that must be produced. There are two arcs between every pair of nodes, one in each direction, representing the changeover time between the corresponding products. A solution to the original problem corresponds to

determining a Hamiltonian path in this graph, i.e., a path going through all the nodes in the graph. The path must start with a particular node (the item being currently produced), but there is no concern about the ending node. Let us call this the “*Fixed-Start Asymmetric Hamiltonian Path*” (FSAHP) problem.

Given the similarity of this problem with the Travelling Salesman Problem, in particular with its asymmetric variants, we considered adapting the methods that have been developed for that problem to the current situation. Throughout this paper we will describe more formally the problem in mathematical programming, explain in detail the metaheuristics that we implemented for solving it, and present the results of applying it to a set of benchmark problems.

2 Problem Description

We are given a graph $G(V, A)$ where V is the set of nodes and A the set of arcs. In the classical Asymmetric Travelling Salesman Problem (ATSP), nodes correspond to cities to be visited and arcs to the distance between them. In our case, each node represents a product to be manufactured, every arc (i, j) has a cost $D(i, j)$ corresponding to the (asymmetric) changeover time between product i and j , and there is a special node v_1 , which must be the first node in the path, and corresponds to the last previously manufactured product (or to the city where the salesperson currently is, the classical problem).

With simple data preprocessing, standard ATSP formulations can be adapted to the current problem, as shown below.

Property 1. Redefine the distance from any node to the first (fixed) node in the path, v_1 , as zero (all other distances remaining unchanged). A minimum Hamiltonian cycle determined with this data defines a path which is an optimal solution to the FSAHP, with the same optimal objective value.

Proof. Let us call the optimal solution to the FSAHP $(p_1, \dots, p_n)$; this is a path, with $p_1 = v_1$, covering all the nodes. This path can be extended into a cycle, without increasing the cost, by adding the arc (p_n, v_1) .

Suppose there is a cycle $(s_1, \dots, s_n, s_1)$, with $s_1 = v_1$, with a smaller objective; then, as the arc (s_n, v_1) has zero length, the path $(s_1, \dots, s_n)$ would have to be shorter than $(p_1, \dots, p_n)$. But in this case $(s_1, \dots, s_n)$ would be a better solution to the FSAHP than $(p_1, \dots, p_n)$, contradicting the assumption.

2.1 Formulation in Mathematical Programming

There are many formulations for the ATSP, and their study is an active field in mathematical programming. For the purposes of this paper, we will restrict to the most common one, due to [1]:

$$\text{minimise } \sum_{i=1}^n \sum_{j=1}^n c_{ij} x_{ij} \quad (1)$$

$$\begin{aligned}
 \sum_{i=1}^n x_{ij} &= 1, \quad j = 1, \dots, n \\
 \sum_{j=1}^n x_{ij} &= 1, \quad i = 1, \dots, n \\
 (n-1)x_{ij} + u_i - u_j &\leq n-2, \quad i, j = 2, \dots, n \\
 x_{ij} &\in \{0, 1\}, \quad i, j = 1, \dots, n, \quad u_i \in \mathbb{R} \quad i = 1, \dots, n
 \end{aligned}$$

The optimal cycle is the set of arcs (i, j) such that $x_{ij} = 1$. The solution to the FSAHP is the n -node path starting with v_1 in this cycle.

3 Basic Heuristics and Local Search

The most straightforward way for solving the Fixed-Start Asymmetric Hamiltonian Path with heuristics and metaheuristics is to apply the transformation on the data proposed in Section 2, and solve an Asymmetric Travelling Salesman. Then, the solution to the original problem is obtained by selecting the n -node path starting with v_1 in the ATSP's solution.

The characteristics of the path problem could be exploited for devising more adapted neighbourhoods, but it turns out that the performance degrades in most of the studied instances, possibly due to the losing symmetry properties.

3.1 Construction

Simple construction heuristics for the ATSP are based on equivalent heuristics for the symmetric TSP (nicely described e.g. in [2]). As for the metaheuristics described in this paper, the initial solution is constructed based on a random permutation of $\{1, \dots, n\}$.

3.2 Improvement

The most common improvement methods for problems related to the TSP are based on *exchange heuristics*: remove k edges, breaking the cycle tour into k paths; then reconnect those paths into a different cycle [3,4]. For the symmetric TSP, the most commonly used neighbourhood is 2-Exchange: remove two non-consecutive edges, and add two other edges, as shown in Figure 1.

As for the ATSP, there are no 2-Exchange moves that keep path orientation, and hence they are not usually employed [5]. The most commonly used moves are 3-Exchange, keeping path orientation, as shown in Figure 2.

For implementing local search based in this neighbourhood in an efficient way, moves that are known not to lead to improvements should be avoided. For this purpose, the list of the neighbours of a given vertex, sorted by distance, is searched only up to a certain point.

Let us first recall what is commonly done with the (symmetric) TSP. Consider a tour represented by $p = (p_1, p_2, \dots, p_n)$, and let us denote the last element of p as either p_n or p_0 . Each edge (p_{i-1}, p_i) , for $i = 1, \dots, n$, is examined for improving exchanges, through removing it and another edge (p_{j-1}, p_j) , and adding two different edges, in such a way that a new tour is formed (p_j must be separated from p_i by at least two nodes). A new tour is constructed by adding edges $\{p_{i-1}, p_{j-1}\}$ and $\{p_i, p_j\}$.

Property 2. For a given i , improving moves cannot be missed if j is restricted to:

1. nodes connected to p_{i-1} such that their distance to p_{j-1} is smaller than $D(p_{i-1}, p_i)$;
2. nodes connected to p_i such that their distance to p_i is smaller than $D(p_{i-1}, p_i)$.

Proof. Let $p_{i-1}, p_i, p_{j-1}, p_j$ be represented by a, b, c, d , respectively, as in Figure 1. In an improving move there must be $D(a, c) + D(b, d) < D(a, b) + D(c, d)$, implying that either $D(a, d) < D(a, b)$ or $D(c, b) < D(c, d)$, or both. Hence, in an improvement, at least one of the added edges must be smaller than at least one of the removed edges. The case of an added edge being smaller than $\{a, b\}$ is examined by considering all edges $\{a, c\}$ such that $D(a, c) < D(a, b)$, and all edges $\{b, d\}$ such that $D(b, d) < D(a, b)$. The remaining potential improvement case corresponds to having the edge $\{c, d\}$ larger than either $\{a, c\}$ or $\{b, d\}$; but this possibility is examined for i such that $c = p_{i-1}$ and $d = p_i$.

Let us now go back to the ATSP problem and the 3-Exchange neighbourhood. Consider a tour represented by $p = (p_1, p_2, \dots, p_n)$. Each arc (p_{i-1}, p_i) , for $i = 1, \dots, n$, is examined for improving exchanges, through removing it and other two arcs (p_{j-1}, p_j) and (p_{k-1}, p_k) . A new tour is constructed by adding arcs (p_{i-1}, p_j) , (p_{j-1}, p_k) , and (p_{k-1}, p_i) .

Property 3. For a given i , improving moves cannot be missed if j and k are restricted as follows:

1. j is restricted to nodes outgoing from p_{i-1} such that their distance from p_{i-1} is smaller than $D(p_{i-1}, p_i)$; furthermore, in this case k is restricted to nodes outgoing from p_{j-1} such that distance $D(p_{i-1}, p_j) + D(p_{j-1}, p_k)$ is smaller than $D(p_{i-1}, p_i) + D(p_{j-1}, p_j)$, and p_k is not in the path from p_i to p_{j-1} .
2. $k - 1$ is restricted to nodes incoming into p_i such that their distance to p_i is smaller than $D(p_{i-1}, p_i)$; furthermore, in this case j is restricted to nodes incoming into p_k such that distance $D(p_{k-1}, p_i) + D(p_{j-1}, p_k)$ is smaller than $D(p_{i-1}, p_i) + D(p_{k-1}, p_k)$, and p_j is not in the path from p_k to p_{i-1} .

Proof. Let $p_{i-1}, p_i, p_{j-1}, p_j, p_{k-1}, p_k$ be represented by a, b, c, d, e, f , respectively, as in Figure 2. In an improving move there must be $D(a, d) + D(c, f) + D(e, b) < D(a, b) + D(c, d) + D(e, f)$, implying that at least one of the added arcs must be smaller than at least one of the removed ones.

Let us consider an improving move for which either $D(a, d) + D(c, f) > D(a, b) + D(c, d)$, or $D(a, d) > D(a, b)$. In the former case, there must be

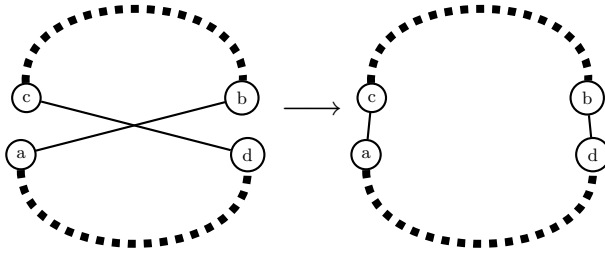


Fig. 1. Single 2-Exchange possibility for the (symmetric) TSP. Edges $\{a, b\}$, $\{c, d\}$ are removed, and replaced by $\{a, c\}$, $\{b, d\}$.

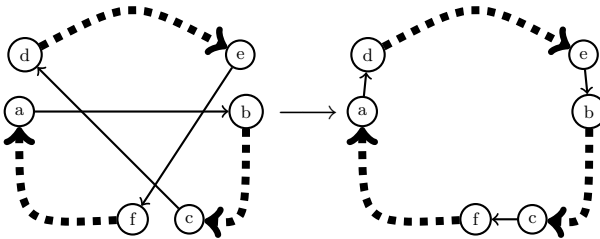


Fig. 2. Single 3-Exchange possibility without path inversions for the ATSP. Arcs (a, b) , (c, d) , (e, f) are replaced by (a, d) , (e, b) , (c, f) .

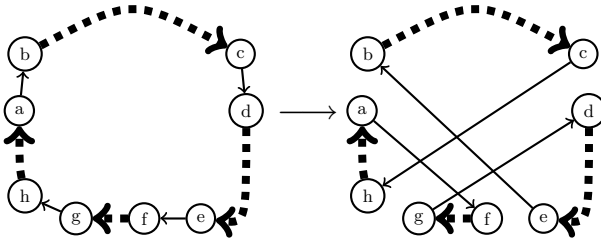


Fig. 3. A 4-Exchange (double bridge) movement without path inversions for the ATSP, as implemented in iterated local search

$D(e, b) < D(e, f)$, and this is tackled in the main i cycle, for $i : p_{i-1} = e$. As for the latter case, there must be $D(c, f) + D(e, b) < D(c, d) + D(e, f)$; thus, either $D(c, f) < D(c, d)$, or $D(e, b) < D(e, f)$, or both. But this situation is tackled for $i : p_{i-1} = c$ or $i : p_{i-1} = e$, respectively.

3.3 Implementation

In our implementation, indices for the outer cycle (i) are searched in random order. Indices j and k are search by increasing distance to nodes p_{i-1} and p_i , until reaching the limits defined by Property 3. Improvements are accepted in an *first-improve* manner, i.e., an improving movement is immediately accepted. The initial solution is a random permutation of $\{1, \dots, n\}$.

3.4 Improved Heuristics

Random-start local search: in this metaheuristics, the following steps are repeated until reaching a stopping criterion (in our implementation, exceeding the limit CPU time):

1. create a random solution;
2. improve it until reaching a local optimum;
3. possibly, update the best solution found so far.

Iterated local search: for this metaheuristics, after reaching a local optimum a deep modification on the solution structure is introduced; the solution thus obtained is then improved until reaching another local optimum, and the whole process is repeated until reaching the stopping criterion.

The deep modification made at each iteration is a 4-Exchange movement, as depicted in Figure 3. This is usually called a “*double bridge*” movement. Our implementation of iterated local search consists of obtaining a random starting solution, and then repeating the following steps:

1. improve the solution until reaching a local optimum;
2. possibly, update the best solution found so far;
3. randomly select 4 arcs in the solution; exchange them with 4 different arcs, in such a way that a tour (with no path inversions) is formed.

4 Results

The metaheuristics proposed in this paper were compared to a mixed-integer programming (MIP) solver, through an experiment with a set of standard benchmarks instances. These correspond a modification of the ATSP instances available in the TSPLIB [6]; the starting node v_1 is the first city in the instance, and, for tackling the path problem, the distances from any other node to v_1 are redefined as zero (as described in Section 2). The experiment was run in a computer with a Quad-Core Intel Xeon, 2.66 GHz processor, running the Mac OS X operating system version 10.6.3; only one CPU was allocated to this experiment. The MIP solver used is GUROBI [7], one of the leading commercial solvers. Metaheuristics were implemented in the Python programming language, version 2.6.1; this is considerably slower than the compiled, executable code of GUROBI. Hence, results are not truly comparable; however, they still allow drawing many interesting conclusions. In all the experiments, the CPU time for an observation of a method solving an instance was limited to about 300 seconds; as for the metaheuristics, the results correspond to the minimum, average, and maximum of 10 independent observations. The results are presented in table 1.

The first interesting conclusion is that a state-of-the-art MIP solver can reach the optimum for many of the benchmark instances (those for which the lower bound obtained is identical to the upper bound); this is an enormous progress with respect to some years ago. In these cases, both metaheuristics could also

Table 1. Results obtained using multi-start local search, iterated Local search, and the lower and upper bounds obtained by the MIP solver *GUROBI*, for a CPU limit of 300 seconds (Instances **dc563**, **dc895**, **dc932** were allowed only one descent, as it takes more than 300 seconds.)

Instance	Multi-start local search			Iterated local search			GUROBI	
	minimum	average	maximum	minimum	average	maximum	LB	UB
atex1	1564	1564	1564	1564	1564	1564	1564	1564
atex3	2342	2342	2342	2342	2342	2342	2342	2342
atex4	2681	2681	2681	2681	2681	2681	2681	2681
atex5	4659	4663.8	4669	4659	4670.8	4747	4595	4659
atex8	41531	41763	41960	41299	41598.8	41900	1027	∞
big702	78933	79081.4	79316	78492	78847.4	79518	−∞	∞
br17	27	27	27	27	27	27	27	27
code198	4541	4541	4541	4541	4541	4541	4541	4541
code253	106957	106957	106957	106957	107032	107333	105716	∞
dc112	10916	10919.3	10922	10914	10916.7	10919	10860	10968
dc126	120725	120770	120827	120709	120754	120808	119702	126506
dc134	5543	5544.6	5547	5539	5540.8	5542	5529	∞
dc176	8402	8406.3	8410	8400	8403.3	8409	8356	∞
dc188	9977	9979.9	9986	9974	9979.8	9988	9911	∞
dc563*	25880	25880	25880	25880	25880	25880	25687	∞
dc849	37496	37501.7	37506	37488	37498.6	37504	−∞	∞
dc895*	106963	106963	106963	106963	106963	106963	−∞	∞
dc932*	478316	478316	478316	478316	478316	478316	−∞	∞
ft53	6099	6099	6099	6099	6099	6099	6099	6099
ft70	37230	37231.2	37234	37230	37230.4	37234	37228	37230
ftv100	1743	1746.5	1747	1743	1744.7	1747	1743	1743
ftv110	1908	1910.6	1914	1908	1912.3	1917	1900	1908
ftv120	2074	2078.2	2081	2074	2074.5	2077	2074	2074
ftv130	2240	2250.2	2262	2240	2242.7	2250	2240	2240
ftv140	2358	2364.4	2375	2358	2360.1	2366	2356	2356
ftv150	2547	2554.5	2563	2547	2548.1	2550	2547	2547
ftv160	2600	2605.5	2616	2600	2603.1	2605	2600	2600
ftv170	2690	2701.7	2717	2689	2691.4	2694	2668	2713
ftv33	1223	1223	1223	1223	1223	1223	1223	1223
ftv35	1363	1363	1363	1363	1363	1363	1363	1363
ftv38	1438	1438	1438	1438	1438	1438	1438	1438
ftv44	1535	1535	1535	1535	1535	1535	1535	1535
ftv47	1689	1689	1689	1689	1689	1689	1689	1689
ftv55	1539	1539	1539	1539	1539	1539	1539	1539
ftv64	1726	1726	1726	1726	1726	1726	1726	1726
ftv70	1881	1881	1881	1881	1881	1881	1881	1881
ftv90	1538	1538	1538	1538	1538	1538	1538	1538
kro124p	35584	35584	35584	35584	35584	35584	35581	35584
p43	589	589	589	589	589	589	549	589
rbg323	1308	1308	1308	1308	1308	1308	1308	1308
rbg358	1143	1143	1143	1143	1143	1143	1143	1143
rbg403	2450	2450	2450	2450	2450	2450	2450	∞
rbg443	2710	2710	2710	2710	2711.7	2719	2710	∞
ry48p	13870	13870	13870	13870	13870	13870	13869	13870
td100.1	267047	267047	267047	267047	267047	267047	267047	267058
td1000.20	1241220	1241230	1241230	1241220	1241230	1241230	−∞	∞
td316.10	688929	688929	688929	688929	688929	688929	688929	688929

find systematically the optimum, except for instances of the **ftv** series. For these instances and **atex5**, the result of the MIP solver is better than the average solution of each metaheuristics; for all the other instances, both metaheuristics are better.

A very interesting result was obtained for instances **rbg403** and **rbg443**; indeed, even though no feasible solution was found by the MIP solver in the CPU

time allowed, the best solution found by metaheuristics can be proven optimal, as its objective value equals the MIP lower bound.

As for the comparison between the two metaheuristics proposed, iterated local search is at least as good as multi-start local search for most instances, being strictly better for many of them; the slight increase in complexity seems, hence, to be worthy.

5 Conclusions

In this paper we describe a variant of the Asymmetric Hamiltonian Path Problem, with applications in scheduling. We present a mathematical programming formulation, and simple approximative methods for solving it. The metaheuristics are random-start local search and iterated local search; both of them provided very good results, with a slight advantage to the latter.

For easy problems a mixed-integer programming solver could find the optimum in a relatively small time; for larger, more difficult problems, the approximative methods could find better solutions in the CPU time allowed.

Improvements on the metaheuristics are expected if “don’t look bits” are used, in order to keep track of cities for which search could be skipped. Another possible improvement concerns limiting the number of neighbours of each city that are allowed to be explored for exchanges. Both of these modifications may provide a considerable speedup, at the cost of, possibly, losing local optimality.

Acknowledgments. This research was supported in part by FCT – Fundação para a Ciência e a Tecnologia (Project **PTDC/GES/73801/2006) and by the European project CIVITAS-ELAN, under Framework Programme 7.

Our special thanks to Prof. Nelma Moreira for proof reading this manuscript.

References

1. Miller, C.E., Tucker, A.W., Zemlin, R.A.: Integer programming formulation of traveling salesman problems. *J. ACM* 7(4), 326–329 (1960)
2. Johnson, D., McGeoch, L.: Local search in combinatorial optimization. In: Aarts, E., Lenstra, J.K. (eds.) *Local Search in Combinatorial Optimization*. John Wiley & Sons, Inc., New York (1997)
3. Croes, G.A.: A method for solving traveling-salesman problems. *Operations Research* 6, 791–812 (1958)
4. Flood, M.M.: The traveling-salesman problem. *Operations Research* 4, 61–75 (1956)
5. Johnson, D.S., Gutin, G., McGeoch, L.A., Yeo, A., Zhang, W., Zverovitch, A.: Experimental analysis of heuristics for the atsp. In: Gutin, G., Punnen, A.P. (eds.) *The Traveling Salesman Problem and Its Variations*. Combinatorial Optimization, vol. 12. Kluwer Academic Publishers, Boston (2002)
6. Bixby, B., Reinelt, G.: TSPLIB – A library of travelling salesman and related problem instances. Internet repository (1995), <http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/>
7. Gurobi Optimization, Inc.: Gurobi Optimizer Reference Manual, Version 2.0 (2010), <http://www.gurobi.com>

Adaptive Intelligence Applied to Numerical Optimisation

Kalin Penev¹ and Anton Ruzhekov²

¹ Southampton Solent University, UK
`Kalin.Penev@solent.ac.uk`

² Technical University of Sofia, Bulgaria
`a_ruzhekov@tu-sofia.bg`

Abstract. The article presents modification strategies' theoretical comparison and experimental results achieved by adaptive heuristics applied to numerical optimisation of several non-constraint test functions. The aims of the study are to identify and compare how adaptive search heuristics behave within heterogeneous search space without retuning of the search parameters. The achieved results are summarised and analysed, which could be used for comparison to other methods and further investigation.

Keywords: Free Search, optimisation, adaptive search heuristics, Genetic Algorithm, Particle Swarm Optimisation, Differential Evolution.

1 Introduction

A previous study [10] compares Free Search (FS) [11] Particle Swarm Optimisation (PSO) [4], and Differential Evolution (DE) [14] on several heterogeneous numerical problems. This article presents another investigation, which compares modification strategies of real-value coded Genetic Algorithm BLX-a (GA BLX-a) [6], PSO [4], DE [14] and FS [11]. In order to assess their ability for adaptation these algorithms are applied without changes of their parameters to several test problems. The aim is to compare how these algorithms behave within heterogeneous search space without retuning of the search parameters.

2 Genetic Algorithm

Genetic Algorithms are computational models inspired by the concept about natural selection and evolution of the biological species described by Charles Darwin in "The Origin of Species". Natural evolution can be considered as a kind of search process. Therefore this concept is recognised as valuable in the domain of heuristics optimisation and search methods. A computational implementation and application of Genetic Algorithms are proposed by Holland [9]. Genetic Algorithms are different from other optimisation and search processes in several ways: (1) GAs work with a coding of the parameter set, not the parameters themselves; (2) GAs search from a population of points, not from a single

point; (3) GAs use payoff (objective function) information, not derivatives or other auxiliary knowledge; (4) GAs use probabilistic transition rules, not deterministic rules [7]. A GAs major event is modification. It involves selection of parents, recombination between them, mutation and evaluation. For this study a Blend crossover modification strategy called BLX- α [6] is selected. For BLX- α modification strategy, the offspring is a random location within the area determined by selected parents and extended with a blend interval α . The mathematical description of BLX- α modification strategy is presented at equation 1:

$$X_{offspring} = X_{p1} - \alpha + (X_{p2} - X_{p1} + 2\alpha) * random(0, 1) \quad (1)$$

where X_{p2} and X_{p1} are selected parents, $X_{p2} > X_{p1}$, α is a blend around the selected parents, $random(0, 1)$ generates a random value between 0 and 1.

An extension of the space, between selected parents, increases the chances of the algorithm to reach an appropriate solution if it is near to the area determined by the parents. Variation of the blend α can be used for tuning of the search process convergence and divergence. Therefore, the concept for extension of the space for modification by a blend α is considered as valuable for improvement of the performance of the search process. For the purposes of the investigation the GA BLX- α is modified and implemented with a variable blend α . A low level of blend α , benefits convergence to the optimal solution and improves effectiveness of the search process by decreasing the number of generations necessary to attain the optimum. However, it takes a risk of being trapped in local sub-optima. A high level of blend α benefits diversification of the population and decreases the probability for trapping in non-optimal areas, which improves the algorithm robustness. The optimisation process trapping, in a non-optimal area, cannot be resolved by variation of the blend value, due to a lack of knowledge how to tune the blend, abstracted from the current population. This problem can be a subject of further research. In summary real value GA BLX- α implicitly determines search space, as promising, with non-zero probability for generation of an offspring, and non-promising, with zero probability for generation of an offspring. For uni-modal problems with one optimal solution this determinism is excellent and leads to quick convergence to the appropriate solution.

However, for multi-modal problems with many, local, sub-optimal solutions this determination restricts the chances of the search process to reach an appropriate solution if it is outside of the area considered as promising from the current population. It often leads to trapping in a non-optimal solution.

3 Particle Swarm Optimisation

PSO can be classified as a population-based, evolutionary computational paradigm. [4]. It has been compared to Genetic Algorithms [1,5] for efficiently finding optimal or near-optimal solutions in large search spaces. PSO is different from other evolutionary computational methods. It attempts to model a social behaviour of a group of individuals [1,13]. In PSO each particle is defined as a potential solution to a problem in multi-dimensional space. A particle i position is represented as:

$$X_i = (x_{i1}, x_{i2}, \dots, x_{id}) \quad (2)$$

where $i \in (1, n)$, n is population size (number of individuals), d is number of dimensions of the search space. Each particle maintains a memory of its previous best position:

$$P_i = (p_{i1}, p_{i2}, \dots, p_{id}) \quad (3)$$

The Particle Swarm Optimisation consists a concept for particle velocity. The velocity along each dimension is represented as:

$$V_i = (v_{i1}, v_{i2}, \dots, v_{id}) \quad (4)$$

At each iteration, the best fitness vector is memorised and denoted as g .

$$g = \underset{i=1}{M}^n (P_i) \quad (5)$$

The particles' best achievement is denoted as vector P_i . The best achievement for all population is denoted as vector g . The current position of the particle X_i , the best particles' achievement P_i and the best achievement for all population g are used for generation of the velocity vector v for each particle (equation 6). That velocity v is then used to compute a new position for the particle (equation 7). The portion of the adjustments to the velocity influenced by the individual's previous best position P_i is considered as an individual cognition component. The portion influenced by the best of the population is a social component [4]. With the addition of the inertia factor, w [13] the particles are manipulated according to the following equations:

$$v_{id} = w * v_{id} + n_1 * random(0, 1) * (P_{id} - x_{id}) + n_2 * random(0, 1) * (g_d - x_{id}) \quad (6)$$

$$x_{id} = x_{id} + v_{id} \quad (7)$$

Where the constants n_1 and n_2 determine the relative influence of the social and cognitive components, and are usually both set the same to give each component equal weight as the cognitive and social learning rate. n_1 is defined as the individual learning factor and n_2 is defined as the social learning factor. One of the advantages of PSO is that there are few parameters to adjust. One version, with slight variations, works well in a wide variety of applications.

The inertia factor influences PSO positively. Large inertia factor facilitates global exploration and searching new areas, while small inertia factor tends to facilitate local exploration and fine-tunes the current search area [5].

4 Differential Evolution

Differential Evolution is proposed by Price and Storn [12,14]. It starts with a stochastic selection of an initial set of solutions called design vectors. The value

of an objective function, which corresponds to each individual of the population, is a measure of that individual's fitness as an optimum. Then, guided by the principle of survival of the fittest, the initial population of vectors is transformed, generation-by-generation, into a solution vector. DE selects for manipulation target, donor and differential vectors. Therefore the minimal number of vectors in one population has to be more than four. For modification strategies, which use four differential vectors the minimal population size is seven. The current target and the corresponding new trial vector (individual) in each generation are subject of competitions to determine the composition of the next generation. The new trial vector is generated in several steps as follows: (1) selection of a randomly chosen donor vector from the population different from the current target vector; (2) selection of other (two or four) randomly chosen vectors (so called differential vectors), different from the donor, different from the current target vector and different from each other; (3) calculation of a difference between differential vectors and scaling it by multiplication with a constant called differential factor; (4) adding the difference to the donor vector, which produces a new vector; (5) crossover between the current target vector and the new vector so that the trial vector inherits parameters from both of them. If the trial vector is better than the current target vector, then the trial vector replaces the target vector in the next generation. In all, three factors control evolution under DE: the population size; the scaling weight applied to the random differential (noted as F); and the constant that mediates the number of parameters in the crossover operation. They describe DE as a heuristic approach for optimising non-linear and non-differentiable functions within continuous space [14]. Let us denote the target vector - X_k , the differential vectors - X_i and X_j , and the differential factor (weight) - F . Every pair of vectors (X_i, X_j) in the primary array defines a differential vector $X_i - X_j$.

When these two vectors are chosen randomly, their weighted difference is used to perturb another vector in the primary array, X'_k :

$$X'_k = X_k + F(X_i - X_j) \quad (8)$$

F scales the difference achieved from $X_i - X_j$. An effective variation of this scheme involves keeping track of the best vector noted as X^* . This can be combined with X_k and then perturbed, producing:

$$X'_k = X_k + F(X^* - X_k) + F(X_i - X_j) \quad (9)$$

Storn proposes several modification strategies for calculation of a new individual as follows:

$$(1) X'_k = X_k + F(X_i - X_j) \quad (10)$$

$$(2) X'_k = X^* + F(X_i - X_j), \quad (11)$$

$$(3) X'_k = X_k + F(X^* - X_k) + F(X_i - X_j), \quad (12)$$

$$(4) \ X'_k = X * + F(X_i - X_j + X_n - X_m), \quad (13)$$

$$(5) \ X'_k = X_k + F(X * - X_k + X_n - X_m), \quad (14)$$

where X_k is a donor vector, X'_k is mutated donor, X^* is the best vector for current population, X_i , X_j , X_n and X_m are differential vectors, F is differential factor. These strategies can be applied to all the variables, to part of the variables or to one variable of the donor vector. Comparison between modification strategies of DE and PSO suggests that they are very similar. However, these strategies are grounded on different concepts therefore the algorithms behaviour and their results are different.

From another point of view mutation in DE is, in fact, a calculation of the sum between the donor vector and the differential of two other or four other vectors [14]. Comparison of this operation with the BLX-a real-coded crossover [6] can identify similarity between them. In the next step each primary array vector X_k is targeted for recombination with X'_k to produce a trial vector X_t . Thus the trial vector is the child of two parents, a noisy random vector and the primary array vector against which it must compete. Once a new trial solution has been generated, selection determines which among them will survive into the next generation. Each child X_t is pitted against its parent X_k in the primary array. Only the fitter of the two is then allowed to advance into the next generation.

5 Free Search

Free Search is real value adaptive heuristic method inspired by animals behaviour in nature. The search process is organised in exploration walks, which differs from classical iterations [11]. FS modification strategy is described as follow: $Xmin_i$ and $Xmax_i$ denote the search space borders, m is a population size, $j = 1, \dots, m$, $k = 1, \dots, m$, n is a number of dimensions, $i = 1, \dots, n$. T is step limit per walk. t is current step. R_{ji} is a variable neighbouring space $R_{ji} \in [Rmin, Rmax]$. The algorithm requires definition of search space borders $[Xmin_i, Xmax_i]$, population size m , limit for number of explorations G , limit for number of steps per exploration T , minimal and maximal values for the neighbour space $[Rmin, Rmax]$. The maximal neighbour space guarantee coverage of the whole search space from one animal. The minimal neighbour space guarantee desired granularity of the coverage from one animal. $Rmin$ and $Rmax$ are absolute values. An appropriate definition of these values supports successful performance across variety of problems without additional external adjustments [11]. A determination of the neighbour space to concrete value for particular problem can lead to slightly better performance on this problem but aggravates the performance on other problems, which is in line with the existing general assessment of the performance of the optimisation algorithms [15].

The exploration walk in FS generates coordinates of a new location x_{tji} as:

$$x_{tji} = x_{0ji} - \Delta x_{tji} + 2 * \Delta x_{tji} * random_{tji}(0, 1). \quad (15)$$

The modification strategy is:

$$\Delta x_{tji} = R_{ji} * (X_{maxi} - X_{mini}) * random_{tji}(0, 1) \quad (16)$$

Where $i = l$ for uni-dimensional step, $i = 1, \dots, n$ for multi-dimensional step. T is step limit per walk. t is current step, $t = 1, \dots, T$. R_{ji} indicates a neighbour space size for animal j within dimension i . $random_{tji}(0, 1)$ randomises the steps within defined neighbour space. The modification strategy is independent from a current or the best achievements and allows nonzero probability for access to any location of the search space and highly encourages escaping from trapping in local sub-optima.[11]

6 Test Problems

For all experiments the aim is to find the maximum therefore the test functions are transformed in relevant manner. All test problem are in 2 dimensional variant.

Step test function - This test function is proposed by De Jung [3]. It introduces plateaus to the topology. Maximal are all locations, which belong to the plateau $x_i \in [2.0, 2.5)$ and the maximum for 2 dimensions is $f_{max} = 4$. Maximise:

$$f(x_i) = \sum_{i=1}^n [x_i], \text{ where } x_i \in [-2.5, 2.5]. \quad (17)$$

Step sphere test function - It introduces also plateaus to the topology, and also excludes a local correlation of the space. [2] Maximal are all locations, which belong to the plateau $x_i \in [-0.5, 0.5)$. The maximum is $f_{max} = 10$. Maximise:

$$f(x_i) = 10 - \sum_{i=1}^n [x_i + 0.5]^2, \text{ where } x_i \in [-2.5, 2.5]. \quad (18)$$

Michalewics test function - The Michalewics test function is described in the domain of Kyoto University [8]

$$f(x_1, x_2) = \sum_{i=1}^2 \sin(x_i) (\sin(ix_i^2/\pi))^{2m} \quad (19)$$

where $i = 2$, $m = 10$, $x_i \in [0.0, 3.0]$. For two dimension maximum is $f(x_1, x_2) = 1.8013$.

Five hills test function - The Five hills test function is designed for this investigation based on the equation 20 below, where $x_i \in [-10.0, 10.0]$ and $i = 2$.

$$\begin{aligned} f(x_1, x_2) = & 9.4/(1 + 0.05 * ((-x_1)^2 + (-x_2)^2)) + \\ & 9.5/(1 + 1.7 * ((7 - x_1)^2 + (7 - x_2)^2)) + \\ & 9.6/(1 + 1.7 * ((7 + x_1)^2 + (7 + x_2)^2)) + \\ & 9.7/(1 + 1.7 * ((7 - x_1)^2 + (7 + x_2)^2)) + \\ & 10.0/(1 + 1.7 * ((7 + x_1)^2 + (7 - x_2)^2)); \end{aligned} \quad (20)$$

7 Experimental Results

GA, PSO, DE and FS are applied to the above-mentioned functions as follows – Each algorithm is evaluated four times per test function - (1) start from stochastic initial population with limit 100 iterations, (2) start from stochastic initial population with limit 2000 iterations, (3) start from one initial location with limit 100 iterations, (4) start from one initial location with limit 2000 iterations. The single initial location is defined as: $x_0 = x_{min} + 0.9(x_{max} - x_{min})$. Each evaluation is 320 experiments. Population size is 10 (ten) individuals for all algorithms for all experiments. For GA the bled α varies from 0.5 to 1.5. For DE differential factor F varies from 0.5 to 1.5. For PSO inertia W varies from 0.5 to 1.5. For FS neighbour space R varies from 0.5 to 1.5. As successful are accepted results: for Step test function 4; for Step sphere test function 10; for Michalewics test function higher than 1.80 (The maximum is 1.8013.); for Five hills test function higher than 11.6 (The maximum is 11.666.). The number of the successful results from all experiments is presented in Table 1.

Table 1. Experimental results

Heading level		F1	F2	F3	F4	Overall
FS	R*-100	320	320	224	67	931
FS	R-2000	320	320	320	218	1174
FS	OL*-100	320	320	227	74	941
FS	OL-2000	320	320	320	214	1178
DE	R-100	320	320	318	53	1011
DE	R-2000	320	320	319	59	1018
DE	OL-100	-	-	-	-	-
DE	OL-2000	-	-	-	-	-
PSO	R-100	320	320	130	10	780
PSO	R-2000	320	320	169	54	863
PSO	OL-100	-	-	-	-	-
PSO	OL-2000	-	-	-	-	-
GA	R-100	226	244	0	0	470
GA	R-2000	320	320	8	6	654
GA	OL-100	270	314	0	0	584
GA	OL-2000	320	320	16	2	658

F1- Step, F2- Step sphere, F3- Michalewics, F4- Five hills;

* R indicates stochastic initial population; OL indicates start from one location.

The results presented on Table 1 and Figures 1, 2, 3 and 4 suggest that PSO, FS and DE can solve these four tests within 100 iterations, and within 2000 iterations almost any run leads to successful result. DE and PSO due to their modification strategise cannot start form one location. GA begins effective search after the first mutation and has less success.

The results on Step and Step sphere tests functions suggest that GA, PSO, DE and FS can easily manage with absence of local correlation. On Michalewics test DE demonstrates the highest convergence speed. However on global optimization

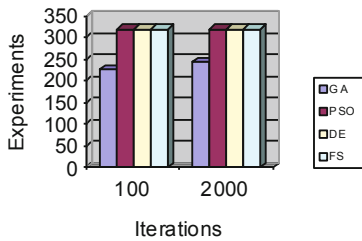


Fig. 1. Step results

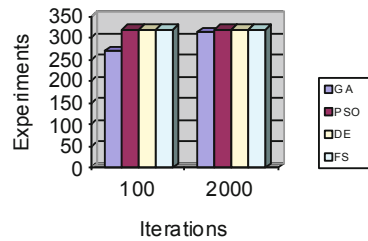


Fig. 2. Step sphere results

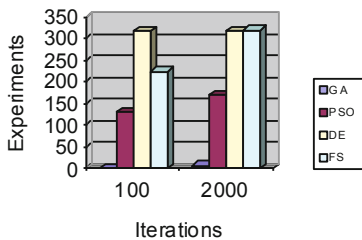


Fig. 3. Michalewics results

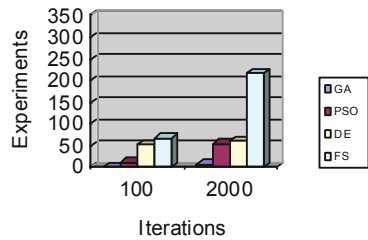


Fig. 4. Five hills result

such as Five hills test the experimental results show that the high convergence speed aggravates adaptation and leads to trapping in local sub optima.

8 Conclusion

The article compares modification strategies of GA BLX- α , PSO, DE and FS and their ability to adapt to four non-constrained tests. Explored algorithms show good capabilities for adaptation to different problems without supervisor's control and without additional adjustment to the concrete problem. This study demonstrates that FS has higher overall performance on explored test. It confirms also Free Search can advance a wide range of disciplines in the efforts to cope with complex problems. Further investigations can focus on replacement strategies comparison and evaluation. A pragmatic area for further research is application to communication tasks such as optimisation of MIMO (Multiple Inputs Multiple Outputs) communication systems.

References

1. Angeline, P.: Evolutionary Optimisation versus Particle Swarm Optimisation: Philosophy and Performance Difference. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, Springer, Heidelberg (1998)
2. Bäck, T., Schwefel, H.-P.: An overview of evolutionary algorithms for parameter optimisation. *Evolutionary Computation* 1(1), 1–23 (1993)

3. De Jong, K.: An Analysis of the Behaviour of a Class of Genetic Adaptive Systems, PhD Thesis, University of Michigan (1975)
4. Eberhart, R., Kennedy, J.: Particle Swarm Optimisation. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
5. Eberhart, R., Shi, Y.: Comparison between Genetic Algorithms and Particle Swarm Optimisation. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447. Springer, Heidelberg (1998)
6. Eshelman, L.J., Schaffer, J.D.: Real-coded genetic algorithms and interval-schemata. In: Foundations of GA, vol. 2, pp. 187–202. Morgan Kaufman Publishers, San Mateo (1993)
7. Goldberg, D.E.: Genetic Algorithms in Search, Optimisation, and Machine Learning. Addison Wesley Longman Inc., Amsterdam (1989) ISBN 0-201-15767-5
8. Hedar, A.R.: Global Optimisation, Kyoto University (2010), http://www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar_files/TestG0_files/Page2376.htm (last visited 02.06.10)
9. Holland, J.: Adaptation In Natural and Artificial Systems. Uni. of Michigan Press, Ann Arbor (1975)
10. Penev, K., Littlefair, G.: Free Search – A Comparative Analysis. Information Sciences Journal 172(1-2), 173–193 (2005)
11. Penev, K.: Free Search of Real Value or How to Make Computers Think. In: Gegov, A. (ed.), UK, April 2008. St. Qu publisher (April 2008) ISBN 978-0955894800
12. Price, K., Storn, R.: Differential Evolution. Dr, Dobb's Journal 22(4), 18–24 (1997)
13. Shi, Y., Eberhart, R.C.: Parameter Selection in Particle Swarm Optimisation. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998)
14. Storn, R., Price, K.: Differential Evolution – A simple and efficient adaptive scheme for global optimisation over continuous spaces, TR-95-012, International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704-1198, Suite 600 (1995)
15. Wolpert, D.H., Macready, W.G.: No Free Lunch Theorems for Optimisation. IEEE Trans. Evolutionary Computation 1(1), 67–82 (1997)

Fed-Batch Cultivation Control Based on Genetic Algorithm PID Controller Tuning

Olympia Roeva¹ and Tsonyo Slavov²

¹ Centre of Biomedical Engineering - BAS, Bulgaria
olympia@cblme.bas.bg

² Technical University - Sofia, Bulgaria
ts_slavov@tu-sofia.bg

Abstract. In this paper a universal discrete PID controller for the control of *E. coli* fed-batch cultivation processes is designed. The controller is used to control feed rate and to maintain glucose concentration at the desired set point. Tuning the PID controller, to achieve good closed-loop system performance, using genetic algorithms is proposed. As a result the optimal PID controller settings are obtained. For a short time the controller sets the control variable and maintains it at the desired set point during the process. Application of the designed controller provides maintaining of the accuracy and efficiency of the system performance.

1 Introduction

A number of processes in the biochemical industry are controlled using PID (proportional-integral-derivative) controllers. Until now commercially available controllers exist only for well established measurement systems as per pH, temperature, stirrer speed, dissolved oxygen etc. The reason for this is highly changing dynamics of most bioprocesses, which is caused by the non-linear growth of the cells, the metabolic changes as well as changes in the overall metabolism. That is, since the PID controller is usually poorly tuned. A higher degree of experience and technology are required for the tuning in a real plant. Tuning a PID controller appears to be conceptually intuitive but can be hard in practice, if complex systems, as cultivation processes are considered. Due to a change of the system parameters, the conventional PID controllers result in sub-optimal corrective actions and hence require retuning. This stimulates the development of tools that can assist engineers to achieve the best overall PID control for the entire operating envelope of a given process. While for control of continuous cultivation processes the controller tuning could be done with traditional methodology, as it is presented in [8], for fed-batch cultivation processes such methodologies could not be applied. For the quality controller tuning optimization methods could be applied, although the tuning procedure is a big challenge for the conventional optimization methods. As an alternative to overcome the controller tuning difficulties various metaheuristics, for example genetic algorithms (GA), could be used [8,5].

This paper focuses on an optimal tuning of universal digital PID controller for control of an *E. coli* fed-batch cultivation process. To achieve good closed-loop system performance GA based controller tuning is proposed. The GA are highly relevant for industrial applications, because they are capable of handling problems with non-linear constraints, multiple objectives, and dynamic components - properties that frequently appear in the real-world problems [8]. Since its introduction and subsequent popularization [4], the GA has been frequently utilized as an alternative optimization tool to the conventional methods [9].

The paper is organized as follows: theoretical background of the GA and of the control algorithm are presented respectively in Section 2 and Section 3. The considered *E. coli* cultivation process is described in Section 4. Controller tuning problem is formulated in Section 5. The results and discussion are presented in Section 6. Conclusion remarks are done in Section 7.

2 Background of the Genetic Algorithms

Genetic algorithms are a class of non-gradient methods. The basic idea of GA is the mechanism of natural selection. Each optimization parameter, x_n , is coded into a gene as for example a real number or string of bits. The corresponding genes for all parameters, $x_1, \dots, x_n$, form a chromosome, which describes each individual. Each individual represents a possible solution, and a set of individuals form a population. In a population, the fittest are selected for mating. Mating is performed by combining genes from different parents by crossover to produce a child. Solutions are also “mutated” by making a small change to a single element of the solution. Finally the children are inserted into the population and the procedure starts over again. The optimization continues until the end-condition is satisfied.

Initial population: A GA starts with a population of strings to be able to generate successive populations of strings afterwards. The initialization is usually done randomly. *Evaluation:* After every generated population, the individuals of the population must be evaluated to be able to distinguish between good and bad individuals. This is done by mapping the objective function to a “fitness function”: a non-negative figure of merit. *Reproduction:* An important aspect is to decide, which individuals should be chosen as parents for the purpose of procreation. With GA, this selection is based on the string fitness: according to the “survival of the fittest” principle. *Recombination:* Once two parents have been selected, the GA combines them to create two new offspring using crossover operator. The role of the crossover operator is to allow the advantageous traits to be spread throughout the population in order that the population as a whole may benefit from this chance discovery [9]. The crossover is the prime distinguishing factor of a GA from other optimization algorithms. *Mutation:* The last operator is the mutation algorithm. The effect of mutation is to reintroduce divergence into a converging population. The biological inspiration behind this operator is the way in which a chance mutation in a natural chromosome can lead to the development of desirable traits giving the individuals advantageous characteristics over its competitors [9].

A pseudo code of a GA is presented as:

$i = 0$	set generation number to zero
initpopulation $P(0)$	initialize a usually random population of individuals
evaluate $P(0)$	evaluate fitness of all initial individuals of population
while (not done) do	test for termination criterion (time, fitness, etc.)
begin	
$i = i + 1$	increase the generation number
select $P(i)$ from $P(i - 1)$	select a sub-population for offspring reproduction
recombine $P(i)$	recombine the genes of selected parents
mutate $P(i)$	perturb the mated population stochastically
evaluate $P(i)$	evaluate its new fitness
end	

3 Background of the Control Algorithm

A PID controller is a generic control algorithm widely used in industrial control systems. The controller parameters used in the calculation must be tuned according to the nature of the system. The standard PID controller calculation (algorithm) involves three separate modes; the proportional (P), the integral (I) and derivative (D). The P mode determines the reaction to the current error, the I mode determines the reaction based on the sum of recent errors, and the D mode determines the reaction based on the rate at which the error has been changing. The weighted sum of these three actions is used to adjust the process via a control element.

In this paper a universal digital PID controller is used due to unsatisfactory performance of control system based on a standard PID controller. A typical structure of a PID control system is shown in Fig. 1. The error signal $e(t)$ is used to generate the P, I, and D modes, with the resulting signals weighted and summed to form the control signal $u(t)$ applied to the plant model. Introducing coefficients b , c and a first-order low pass filter in D mode leads to a negligibly more complex controller, but significantly improves the control system's performance. The coefficient b ($b \leq 1$) is used to weight out the $r(t)$ in P mode of controller and the coefficient c ($c \leq 1$) is used to weight out the $r(t)$ in D mode of the controller. Typically in industrial applications b and c are chosen to be equal to 0 or 1. Using of a first-order low pass filter reduces the influence of measurement noise. In real applications discrete time PID controller is implemented. Many formal techniques for discretization exist [7]. In this paper backward Euler method is used [6]. The mathematical description of discrete-time universal PID controller is:

$$u(k) = u_p(k) + u_i(k) + u_d(k), \quad (1)$$

$$u_p(k) = K_p(br(k) - y(k)), \quad (2)$$

$$u_i(k) = u_i(k-1) + b_{i1}(r(k) - y(k)) + b_{i2}(r(k-1) - y(k-1)), \quad (3)$$

$$u_d(k) = a_d u_d(k-1) + b_d(cr(k) - cr(k-1) - y(k) + y(k-1)), \quad (4)$$

where k is the number of sample, $u(k)$ - control signal, $u_p(k)$, $u_i(k)$ and $u_d(k)$ - proportional, integral and derivative modes of control signal, $r(k)$ - reference signal, $y(k)$ - output signal, K_p - proportional gain, T_i - integral time, T_d - derivative time, T_d/N - time constant of first-order low pass filter, T_0 - sample time, b and c - weighting coefficients, $b_{i1} = K_p \frac{T_0}{T_i}$, $b_{i2} = 0$, $a_d = \frac{T_d}{T_d + NT_0}$, $b_d = K_p \frac{T_d N}{T_d + NT_0}$.

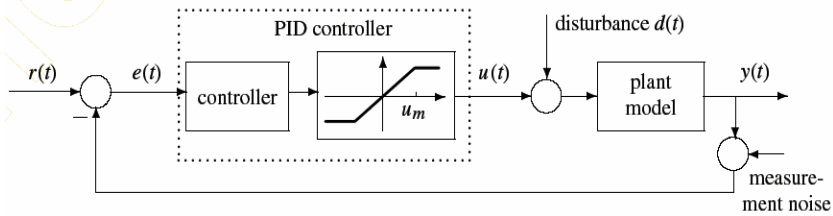


Fig. 1. A typical structure of a PID control system

By tuning the constants (K_p, T_i, T_d, b, c and N) in the PID controller algorithm, the controller can provide control action designed for specific process requirements. Two general tuning methods were proposed by Ziegler and Nichols [11] and have been widely utilized either in the original form or in modified forms. These methods, referred to as "classical" tuning methods, determine the PID parameters using empirical formulae [2,3]. These methods are inapplicable to the considered here non-linear control system. The regarded fed-batch cultivation process can not to be linearized around an equilibrium point of a system. In this case there is no equilibrium point. If a linear approximation is found, the resulting model will be valid only for a small region around the linearization point. The controller tuned using this linear model will work properly only for this limited region. Therefore, it is necessary to use non-classical tuning methods to achieve the best overall PID control for the entire operating envelope of the given system.

4 *E. coli* MC4110 Fed-Batch Cultivation Model

Fed-batch cultivation process of *E. coli* MC4110 is considered. The cultivation conditions and data measurements are discussed in [1]. The mathematical model can be represented by the following dynamic mass balance equations [1]:

$$\frac{dX}{dt} = \mu_{max} \frac{S}{k_S + S} X - \frac{F}{V} X \quad (5)$$

$$\frac{dS}{dt} = -\frac{1}{Y_{S/X}} \mu_{max} \frac{S}{k_S + S} X + \frac{F}{V} (S_{in} - S) + \xi \quad (6)$$

$$\frac{dV}{dt} = F \quad (7)$$

where X is the biomass concentration, [g/l]; S - substrate (glucose) concentration, [g/l]; F - feeding rate, [l/h]; V - bioreactor volume, [l]; S_{in} - substrate concentration in the feeding solution, [g/l]; μ_{max} - maximum value of the specific growth rate, [h^{-1}]; k_S - saturation constant, [g/l]; $Y_{S/X}$ - yield coefficient, [-], ξ - measurement noise. Numerical values of the model parameters used in simulations are according to [1]: $\mu_{max} = 0.55 h^{-1}$, $k_S = 0.01 g/l$, $Y_{S/X} = 0.50$.

5 PID Controller Tuning Using Genetic Algorithm

The simple GA is a powerful tool that is able to converge rapidly to an optimum of many different objective functions. The user has to create a code scheme, a fitness function and implement these into the GA, which mechanisms are easy to implement into a computer program. The optimal value of the PID controller parameters (K_p, T_i, T_d, b, c and N) is to be found using GA.

Initialization of algorithm parameters: The most appropriate GA parameters and operators, based on previous author's investigations on the effects of the different GA parameters on the outcome of the GA [10] are used.

Representation of chromosomes: Representation of chromosomes is a critical part of the GA application. In order to use the GA to identify controller parameters, it is necessary to encode the parameters in accordance with the method of concatenated, multiparameter, mapped, fixed-point coding [4]. Here, a chromosome is a sequence of m - parts each of them with n (encoding precision) genes. In the case of tuning the three controller parameters - K_p, T_i and T_d , the chromosome is a sequence of three parts. In the case of tuning of all the defined parameters - K_p, T_i, T_d, b, c and N , the chromosome is a sequence of six parts. The range of the tuning parameters is considered as follows: $K_p \in [0, 2]$, $T_i \in [0, 1]$, $T_d \in [0, 0.1]$, $b \in [0, 1]$, $c \in [0, 1]$ and $N \in [0.001, 1000]$. After several runs the range for the parameters is specified to: $K_p \in [0.4, 2]$, $T_i \in [0.005, 1]$ and $T_d \in [0.003, 0.1]$.

Following a random initial choice, entire generations of such strings are readily processed in accordance with the basic genetic operators of selection, crossover and mutation. In particular, the selection process ensures that the successive generations of PID controller parameters produced by the GA exhibit progressively improving behavior with respect to some fitness measure.

Objective function: To evaluate the significance of the tuning procedure and controller performance four criteria are used - integrated squared error (I_{ISE}); integrated absolute error (I_{IAE}); integrated time-weighted absolute error (I_{ITAE}) and integrated squared time-weighted error (I_{ISTE}):

$$I_{ISE} = \sum_{k=0}^M e(k)^2, \quad I_{IAE} = \sum_{k=0}^M |e(k)|, \quad I_{ITAE} = \sum_{k=0}^M k e(k), \quad I_{ISTE} = \sum_{k=0}^M k^2 e(k)^2,$$

where the error e is the difference between the set-point and the estimated substrate concentration ($S_{sp} - S$), M - end sample of the cultivation.

Termination criteria: Here the termination criterion is considered to be the maximum number of generations. The chosen maximum number of generations is sufficient for reaching a satisfactory fitness value.

6 Results and Discussion

In the case of cultivation processes control the usual practice is to select PI or PID mode. A P controller reduces error but does not eliminate it, i.e. an offset between the actual and desired value will normally exist. The additional I mode corrects the error that occur between the desired value and the process output. Inclusion of the I mode makes the control system more likely to be oscillatory. Inclusion of the D mode (i.e. selecting PID mode) improves the speed of the responses, and consequently served to suppress the influence of the disturbance more strongly. However, the D mode functions are effective only when the controller parameters are tuned appropriately. Controller tuning is a subjective procedure and is certainly process dependent. For the considered here process the problem is to find a feed rate profile to establish small glucose concentration preventing the accumulation of growth inhibiting metabolites.

Using the considered four objective functions a series of test are performed. To obtain more realistic tests of the controller robustness and of the tuning procedure performance measurement noise is introduced in the simulation - white Gaussian zero mean noise with a variance $0.002 \text{ g}^2/\text{l}^2\text{h}$. For each criterion (in case of noise absence and in case of noise introducing) at least 35 runs of GA are performed. The controller parameters' tuning is performed for two cases: *Case 1* - tuning of the basic PID parameters - K_p, T_i and T_d (parameters b, c and N are defined as constants - $b = c = 1, N = 1000$) and *Case 2* - tuning all the six parameters. The results presented here are mean values of the all runs for the current case. The algorithm produces the same estimations with more than 80% coincidence. Some of the results from the GA application for PID tuning are presented in Table 1. The case with the introduction of noise is shown. This is more real case of the problem decision and the discussion of the corresponding results is more useful.

Table 1. Controller parameters, mean value (with noise)

Case study	K_p	K_i	K_d	b	c	N	I value
1	0.4003	0.9846	0.0030	1.0000	1.0000	1000.0000	$I_{ISE}=16.1639$
2	0.4041	0.9465	0.0030	0.9392	0.8709	778.3171	$I_{ISE}=16.1510$
1	0.4002	0.9853	0.0030	1.0000	1.0000	1000.0000	$I_{IAE}=38.2181$
2	0.4072	0.9356	0.0030	0.8375	0.9283	616.3588	$I_{IAE}=38.1324$
1	0.4002	0.9874	0.0030	1.0000	1.0000	1000.0000	$I_{ITAE}=110.4505$
2	0.4036	0.9353	0.0030	0.9036	0.9060	733.9115	$I_{ITAE}=110.3550$
1	0.4002	0.9783	0.0030	1.0000	1.0000	1000.0000	$I_{ISTE}=755.1833$
2	0.4035	0.9385	0.0030	0.8646	0.9370	537.9127	$I_{ISTE}=754.4549$

The results show that all objective functions are representative and sophisticated controller performance indices. The obtained numerical values of the controller parameters for the four criteria, respectively in *Case 1* and *Case 2* are quite similar. The considered objective functions reflect the performance of the PID controller in a similar way. It could not define the best criterion.

As a result of the GA tuning the optimal PID controller settings are obtained. In Fig. 2 some results of controller and process performance are presented concerning *Case 2* and I_{ITAE} criterion. The obtained results are compared with the results from the controller design of the same cultivation process reported in [1]. In Fig. 2a biomass concentration during the process is displayed. In Fig. 2b and Fig. 2d substrate concentrations and resulting feed rate profiles are presented.

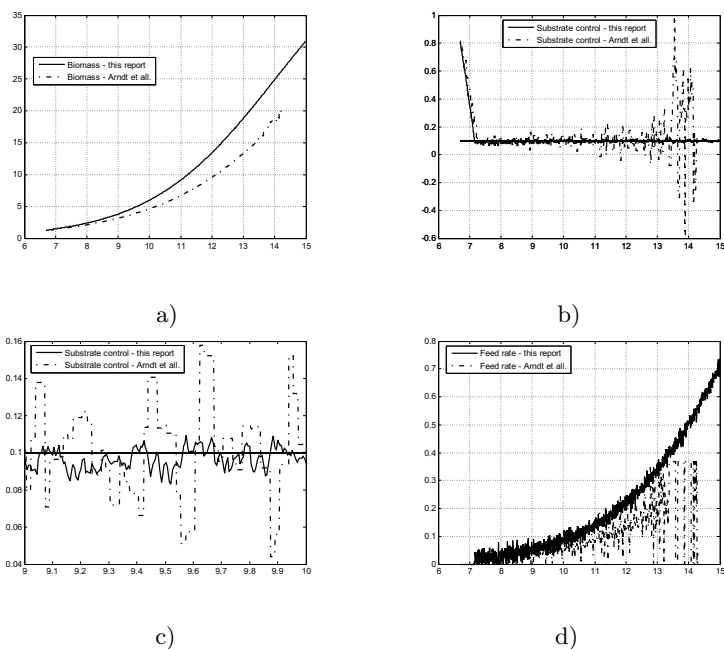


Fig. 2. Controller and process performance

For better visualization in Fig. 2c the substrate concentrations between 9 and 10 h from the cultivation for both studies (this and [1]) are presented. To show the stability of the controller designed here the cultivation process is simulated for a longer time period in comparison with [1]. As it can be seen for a short time the controller sets the control variable and keeps stable the glucose concentration at the set point of 0.1 g/l during the process. The maximum difference reported in [1] is 0.06 g/l and it has occurred in the second half of the process. In parallel, the maximum difference achieved here is 0.028 g/l. Here discussed controller has the better performance than the presented in [1]. The deviation from the set-point is very small for the all time period. The resulting standard deviation and

mean value concerning control variable are: in this report $\rightarrow \sigma_s = 0.0063$ and $m_s = 0.0967$; in [1] $\rightarrow \sigma_s = 0.1513$ and $m_s = 0.1306$.

7 Conclusion

In the article are presented the results of a designed universal digital PID controller. The controller is used to control feed rate and to maintain glucose concentration at the desired set point for an *E. coli* fed-batch cultivation process. GA controller tuning to achieve good closed-loop system performance is proposed. Using four objective functions reflecting the performance of the PID controller, the significance of the tuning procedure is evaluated. As a result, the optimal PID controller settings are obtained. The presented results indicate high quality and better performance of the designed control system. For a short time the controller sets the control variable and maintains it at the desired set point during the cultivation process. It is demonstrated that the GA provide a simple, efficient and accurate approach to PID controllers tuning. Moreover, GA tuning can be regarded as an effective methodology for attaining improved performance of a process.

Acknowledgements. This work is partially supported by National Scientific Fund Grants DMU 02/4 and DID-02-29.

References

1. Arndt, M., Hitzmann, B.: Feed Forward/feedback Control of Glucose Concentration during Cultivation of *Escherichia coli*. In: 8th IFAC Int. Conf. on Comp. Appl. in Biotechn., pp. 425–429 (2001)
2. Astrom, K., Hagglund, T.: PID Controllers, 2nd edn. Instr. Soc. of America (1995)
3. Garipov, E.: PID Controllers. Automatics and Informatics, 3 (2006) (in Bulgarian)
4. Goldberg, D.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley Publishing Company, Massachusetts (1989)
5. Gundogdu, O.: Optimal-tuning of PID Controller Gains using Genetic Algorithms. Journal of Engineering Sciences 11(1), 131–135 (2005)
6. Heath, M.T.: Computing, An Introductory Survey, 2nd edn. McGraw-Hill, New York (2002)
7. Kotsiantis, S., Kanellopoulos, D.: Discretization Techniques: A Recent Survey. GESTS Int. Transact. on Comp. Scien. and Eng. 32(1), 47–58 (2006)
8. Kumar, S.M.G., Jain, R., Anantharaman, N., Dharmalingam, V., Begum, K.M.M.S.: Genetic Algorithm Based PID Controller Tuning for a Model Bioreactor. Indian Chemical Engineer. 50(3), 214–226 (2008)
9. Parker, B.S.: Demonstration of using Genetic Algorithm Learning. Information Systems Teaching Laboratory (1992)
10. Roeva, O.: Improvement of Genetic Algorithm Performance for Identification of Cultivation Process Models. In: 9th WSEAS Int. Conf. on Evol. Comp., pp. 34–39 (2008)
11. Ziegler, J.G., Nichols, N.B.: Optimum Settings for Automatic Controllers. Trans. Amer. Soc. Mech. Eng. 64, 759–768 (1942)

Perspectives of Selfish Behaviour in Mobile Ad Hoc Networks

Marcin Seredynski¹ and Pascal Bouvry²

¹ University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust, 6, rue Coudenhove Kalergi, L-1359, Luxembourg, Luxembourg
`{marcin.seredynski}@uni.lu`

² University of Luxembourg, Faculty of Sciences, Technology and Communication
`{firstname.lastname}@uni.lu`

Abstract. This paper investigates the conditions in which trust-based cooperation on packet forwarding is unlikely to be developed in mobile ad hoc networks. The analysis is performed by combining genetic algorithms and replicator equation dynamics. We demonstrate that in the presence of a large number of unconditionally cooperative nodes a selfish permanent defection forwarding strategy is more successful than a forgiving version of reciprocal tit-for-tat.

Keywords: MANETs, trust management, genetic algorithms, replicator dynamics.

1 Introduction

A civilian wireless mobile ad hoc networks (MANET) consists of a number of wirelessly connected mobile devices (herein referred to as nodes) that are free to move [1]. Such networks operate without a support from any fixed infrastructure as nodes incorporate routing functionality. Most of them rely on batteries, thus temptation to act selfishly by not participating to the packet forwarding duty in order to save battery is high. Nodes might be forced to cooperate on packet forwarding if a distributed *cooperation enforcement* is created (i.e., some nodes start to forward packets only on behalf of those who they find cooperative). In the majority of the cooperation enforcement mechanisms proposed in the literature [2,3,4,5,6] it is assumed that all nodes use the same forwarding approach (herein referred to as forwarding strategy). This assumption might be too strong to hold in civilian MANETs as in such networks nodes belong to different authorities (thus, the use of a given forwarding strategy cannot be enforced). Instead, nodes would rather choose strategies that are most beneficial to them. As demonstrated in the literature [7,8,9,10] defection-tolerant adaptations of a conditionally cooperative strategy called tit-for-tat (TFT) lead to a cooperative network resistant to selfish behaviour. The TFT starts with cooperation (i.e., forwarding packets) and thereafter copies the move of the node

asking for forwarding service. This paper demonstrates that in specific networking conditions, where some nodes have utilitarian preferences, TFT is not the best strategy choice for a node. It is outperformed by a selfish (non-cooperative) strategy. In consequence cooperation enforcement mechanism is not created.

In this work strategies for given networking conditions are discovered using a genetic algorithm (GA) and replicator equation heuristics. This is not a classical optimization problem as an optimal strategy of a node depends on the strategies used by others, thus the fitness landscape is dependent on the strategy frequency. Hence, instead of describing the packet forwarding interactions as a parametric situation a game-theoretical model of MANETs is used. Packet forwarding interactions are modeled as a repeated, sequential two-player game similar to prisoner's dilemma (PD). The paper is structured as follows. Section 2 presents a model of a trust-based packet forwarding. Section 3 explains the evolutionary approach for the analysis of forwarding strategies. Section 4 contains a description of the experimental design and simulation results. The final section summarizes the main conclusions.

2 Model of the Trust-Based Forwarding

Each node uses a strategy that specifies whether a packet received for forwarding should be passed on to the next node. Such a decision is based on the trustworthiness of the node that has originated the packet (source node). Information elements used to derive trustworthiness are represented by two networking events, "packet forwarded" and "packet discarded". A source routing protocol is assumed to be used, thus a list of intermediate nodes is included in the packet header. The information regarding the packet forwarding behaviour of other nodes (referred to as *trust data*) is gathered only by nodes directly participating in the *communication session*. The session involves a source node, forwarders and a destination node. Nodes are equipped with a *watchdog mechanism* (WD) [11], which works as follows: let us assume that node S originates a message to node D via intermediate nodes A and B. The message is next discarded by node B. This event is recorded by the WD mechanism of node A, which next informs node S about the selfish behaviour of B. As a result, the trust system of node S is updated with two events - "packet forwarded by A" and "packet discarded by B", while the trust system of A is updated with the event "packet discarded by B".

In general, node i maintains information about two characteristics of node j : a number of packets forwarded by j ($req\_acc_{j|i}$) and a number of packets discarded by j (only packets originated by i are taken into account). On the basis of these characteristics a *forwarding ratio* of j (ratio of packets forwarded to discarded by j on behalf of i) can be calculated.

The evaluation of trustworthiness of node j by node i is performed in the following way: the sequence of the past actions of j th node concerning packets originated by i th node is divided into three *time frames* (denoted by f_{t-3} (oldest), f_{t-2} and f_{t-1} (newest)). Their sizes (number of forwarding actions taken into account) are denoted by $s_{f_{t-1}}$, $s_{f_{t-2}}$ and $s_{f_{t-3}}$. The sequence of actions

captured by each time frame is next evaluated independently into one of two *trust values* referred to as *cooperative* (*C*) or *selfish* (*S*). These values are calculated on the basis of forwarding ratio and a parameter called *cooperation threshold* (*ct*). If forwarding ratio is greater or equal than the threshold the behaviour of the node captured by the given time frame is classified as cooperative. Otherwise, it is classified as selfish. The goal of the *ct* parameter is to achieve fault-tolerance. Setting its value below 1 allows to evaluate time frames with occasional defections as cooperative ones. Finally, the *forwarding strategy* specifies an action ("F" - forward or "D" - discard) for all possible patterns of trust values of each of the three time frames (see Figure 1). Additionally, each node uses *path rating mechanism* for originating its own packets in order to avoid distrusted nodes as forwarders. If a source node has more than one path available it chooses the one with the best rating, which is calculated as a multiplication of all known forwarding ratios of the nodes belonging to the route.

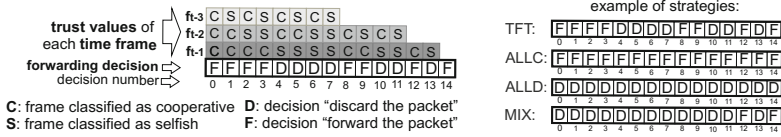


Fig. 1. Trust-based forwarding strategy

3 Evolutionary-Based Analysis of the Network

The packet forwarding interaction is modeled by a non-cooperative, sequential game called *Packet Forwarding* (PF), where players are the nodes in the network, who upon a receipt of a packet for forwarding have two possible choices: forward or discard the packet. Each player prefers to discard a packet (in order to save his battery) and have its own packets forwarded by others. The payoffs that players receive in the PF game are as follows: both receive 3 points for the mutual cooperation, and if player i forwards a packet for j and j does not reciprocate (during duration of the network) then i gets nothing, while j receives 5 points. Players do not receive any pay for mutual defection, i.e., when they mutually discard packets. Moreover, the order of actions is not significant. A total *fitness* received by player i is defined as follows:

$$fitness_i = \sum_{j \in \mathbb{O}_i} \frac{rc_{j|i} \cdot 3 + sc_{j|i} \cdot 5}{rc_{j|i} + sc_{j|i}}, \quad (1)$$

where $\mathbb{O}_i$ denotes a set of all nodes that i interacted with. The $rc_{j|i}$ and $sc_{j|i}$ are given by the following equations:

$$rc_{j|i} = \min(req\_acc_{j|i}, req\_acc_{i|j}), \quad (2)$$

$$sc_{j|i} = \begin{cases} req\_acc_{j|i} - req\_acc_{i|j} & \text{if } req\_acc_{j|i} > req\_acc_{i|j} \\ 0 & \text{if } req\_acc_{j|i} \leq req\_acc_{i|j}. \end{cases} \quad (3)$$

The $rc_{j|i}$ is a number of reciprocated “packet forwarded” actions between the two players, while $sc_{j|i}$ describes a number of cooperative actions of j that were not reciprocated by i .

Two types of nodes are defined: *tester nodes* (TE) and *learner nodes* (LE). TE nodes use fixed strategies. Their goal is to preserve certain properties of the network by using strategies that one could expect to be present in a typical MANET: a TFT (with additional defection-tolerance represented by the ct parameter), an unconditional cooperation strategy representing utilitarian preferences (ALLC) and a selfish “permanent defection” strategy (ALLD). Strategies of LE nodes represent the solution (best response for given conditions), thus they change over time by adapting during the evolutionary process to the networking conditions implicitly defined by all nodes (especially TE nodes). Experimental procedure for the analysis of the network behaviour is defined by algorithm #1. The algorithm can be run either in *evolutionary* or *ecological* mode. In the former new strategies are introduced, while the later results in changes of their frequencies only.

Algorithm #1: analysis of the behaviour of the network

1. set N as a number of nodes in the network, P as a number of TE nodes and M as a number of LE nodes and initialize strategies of LE and TE;
2. simulate MANET as described by algorithm #3;
3. if evolutionary mode is used create a new population of LE strategies replacing the previous one using a GA (with a standard one-point crossover with an uniform bit flip mutation), for the ecological mode use algorithm #2;
4. check the stop condition: if steps 2-3 were repeated predefined number of steps (generations) stop the algorithm. Otherwise go to step 2.

Algorithm #2: new population of strategies by means of replicator equation

1. let x_s^g denote a proportion of nodes in the LE population that use a strategy of type s in generation g ;
2. the new proportion of nodes using strategy s in the subsequent generation (x_s^{g+1}) is given by the following equation:

$$x_s^{g+1} = x_s^g \frac{f_s}{\bar{f}}, \quad (4)$$

where f_s is the average fitness of nodes that used strategy s and $\bar{f}$ is an average fitness of the population of LE nodes.

Algorithm #3: simulation of the network

1. specify i (source node) as $i := 1$ and R as a number of rounds;
2. randomly select node j (destination) and intermediate nodes, forming several possible paths from i th to j th node;

3. for each available path calculate its rating and select the best rated one;
4. let node $i := 1$ originate a packet (initiate a communication session), which is next processed by intermediate nodes according to their forwarding strategies;
5. as soon as the communication session is completed update the trust data;
6. if $i < N$, then choose the next node ($i := i + 1$) and go to step 2. Else go to step 7;
7. if $r < R$, then $r := r + 1$ and go to step 1 (next round). Else stop the simulation and calculate fitness of each player.

4 Experiments

The goal of the experiment was to analyze the conditions favoring selfish forwarding strategies. Several networking settings differing in the composition of the types of strategies were analyzed. The network was composed of 30 TE nodes and 30 LE nodes. Each experiment was repeated 100 times. The performance measures presented in this section describe the mean value of performance of a single node belonging to a given type calculated over all runs of a given experiment. The parameter specifications of the common settings of all experiments are given in Table 1.

Table 1. Parameter specifications of the network and GA

Parameter	Value	Parameter	Value
# of nodes nodes (M)	60 (30 TE, 30 LE)	# of generations	100
simulation time (# of rounds (R))	600	crossover prob.	0.9
	4	mutation prob.	0.001
	0.5	tournament size	2

Two sets of experiments were carried out. In the first one the strategies of LE nodes were discovered for each specified networking condition using the evolutionary mode of algorithm #1. The population of LE nodes was initialized as follows: 15 nodes with TFT strategy, 2 nodes with ALLD strategy and 13 with a randomly generated strategy. The composition of strategies of TE nodes varied, in general, these nodes used either ALLC, TFT or ALLD. Figure 2a demonstrates changes in the forwarding rates of LE nodes in the function of the number of ALLC strategies present in the population of TE nodes. The changes of the throughput rates reached by these nodes are shown in Figure 2b. Three different test cases varying in the proportions of strategies used by TE nodes were considered. In each case a certain number of nodes used the ALLC strategies, while the strategies of the remaining nodes were: equally divided between TFT and ALLD (case 1), only ALLD was used (case 2), only TFT was used (case 3). Detailed values of various performance measures of the selected settings (denoted by $s_1 - s_7$) of the case 1 are shown in Table 2.

One can notice (Figure 2a), that the greater number of ALLC strategies resulted in a higher level of selfishness of the evolved strategies of LE nodes: depending on the test case, whenever at least around 15-18 of TE nodes used the ALLC strategy (25-30% of all nodes), the forwarding ratio of LE nodes dropped

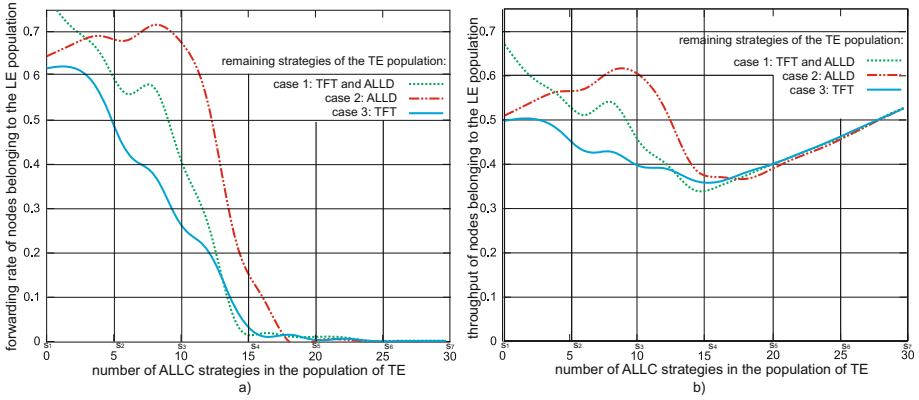


Fig. 2. Performance changes of LE nodes: forwarding rate (a), throughput (b)

Table 2. Results of seven settings of the network (denoted by $s_1 - s_7$). Number of ALLC strategies among TE nodes ranges from 0 to 30, strategies of the remaining TE nodes are equally divided between TFT and ALLD. For each setting the strategies were obtained using the evolutionary mode of algorithm #1.

		number of ALLC strategies among TE nodes						
		0 (s_1)	6 (s_2)	10 (s_3)	16 (s_4)	20 (s_5)	26 (s_6)	30 (s_7)
Learners (LE)	throughput of network	0.56	0.51	0.44	0.38	0.42	0.48	0.53
	throughput of LE	0.67	0.54	0.44	0.38	0.42	0.48	0.53
	σ of throughput of LE	0.04	0.09	0.09	0.011	0.006	0.004	0.004
	# of packets forwarded by LE	772	522	298	6	4	3	2
	average fitness of LE	2.05	1.7	1.36	1.2	1.65	2.52	3.27
	forwarding rate of LE	0.78	0.61	0.40	0.01	0.01	0.01	0.01
	forwarding rate LE vs. LE	0.9	0.67	0.42	0.01	0.01	0.01	0.01
	forwarding rate LE vs. TE	0.64	0.55	0.37	0.01	0.01	0.01	0.01
	forwarding rate LE vs. TFT	0.9	0.66	0.45	0.04	0.05	0.04	-
	forwarding rate LE vs. ALLD	0.18	0.31	0.24	0.01	0.01	0.01	-
Testers (TE)	forwarding rate LE vs. ALLC	-	0.67	0.40	0.01	0.01	0.01	0.01
	dominating strategy in last generation	TFT	MIX	MIX	ALLD	ALLD	ALLD	ALLD
	% runs with the strategy in last gen.	81	28	28	70	80	84	93
	throughput of TE	0.46	0.48	0.45	0.4	0.44	0.49	0.53
	forwarding rate of TE	0.55	0.66	0.69	0.72	0.81	0.93	1.0
	throughput of TFT	0.67	0.59	0.52	0.43	0.45	0.5	-
	throughput of ALLD	0.24	0.32	0.34	0.35	0.4	0.47	-
	throughput of ALLC	-	0.59	0.5	0.42	0.44	0.49	0.52
	# of packets forwarded by TE	390	495	529	556	682	885	1027
	fitness of TFT	2.05	1.62	1.28	0.86	0.98	1.17	-
	fitness of ALLD	0.48	1.53	1.78	1.21	1.69	2.56	-
	fitness of ALLC	-	1.32	0.99	0.72	0.8	0.95	1.05

below 0.1 (meaning that these nodes discarded more than 90% packets received for forwarding). Although, the throughput of these nodes initially decreased to around 0.35, afterwards it started to improve (see Figure 2b). These trends can be seen in each of the three test cases, however surprisingly, greater number of TFT nodes among TE nodes promoted selfishness among LE nodes (see Figure 2a, e.g., case 2 vs. case 3). When it comes to the evolved strategies, two types dominated the cooperative outcomes for LE nodes in $s_1 - s_3$: TFT and a strategy referred further-on to as MIX (for specification of these strategies see

Figure 1). The MIX strategy can be seen as a mixture of TFT (in short term interactions) and ALLD (in long term interactions defined by positions 0–11). It was still quite cooperative towards TFT strategies, with forwarding rate around 0.44. In the remaining settings ($s_4 - s_7$) ALLD strategy was the winner. However, the non-cooperative behaviour of LE nodes did not change significantly the overall performance of the network as the lost contribution of LE nodes was compensated by the generous packet forwarding of unconditionally cooperative nodes (the overall throughput ranged from 0.38 to 0.56).

In the final set of experiments the analysis was performed using the ecological mode of algorithm #1. The initial population of LE nodes was equally divided between the previously found strategies: TFT, ALLD and MIX strategies were assigned to 10 of LE nodes each. Next, networking settings s_1-s_7 were run again using the ecological mode of algorithm #1. The results were as follows: in scenarios s_1 and s_2 the TFT strategy completely dominated the population of LE nodes after on average 19 generations in s_1 and 32 generations in s_2 . In s_3 the TFT was again a clear winner, but it took some time to eliminate the MIX strategy (see Figure 3a). In s_4 all strategies got through to the last generation (although not necessarily in the same runs): on average ALLD was the best one, followed by TFT. In the remaining settings, ALLD was the only surviving strategy (dominated after 17 generations in s_5 , 11 in s_6 , and 13 in s_7). In general the TFT was found to perform better than was suggested in the first set of experiments. Selfish approach towards packet forwarding proved successful only if at least around 20 nodes used the ALLC strategy (one-third of all nodes). The MIX strategy was outperformed in each case (it only managed to survive in some runs of s_4).

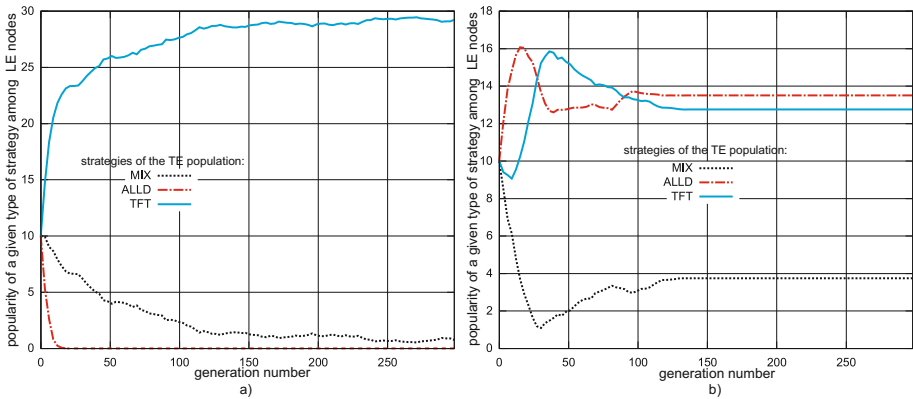


Fig. 3. Changes in the frequencies of strategies: in s_3 (a), in s_4 (b)

5 Conclusion

In this paper we have demonstrated that in the presence of at least one-third of nodes using an unconditionally cooperative strategy a free-riding behavior is the

best choice for network participants. Although the contribution to packet forwarding by selfish nodes is marginal, they obtain similar throughputs as nodes that use reciprocal TFT strategy. The question remains, how nodes could assess the strategies used by other network participants. This is an essential requirement in strategy selection process and constitutes our future work.

Acknowledgments

This work has been partially founded by C08/IS/21 TITAN Project (CORE programme) financed by National Research Fund of Luxembourg.

References

1. Corson, S., Macker, J.: Mobile ad hoc networking (manet): Routing protocol performance issues and evaluation considerations. IETF RFC 2501 (1999)
2. Buchegger, S., Boudec, J.Y.L.: Performance analysis of the confidant protocol. In: Proc. 3rd International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2002), pp. 226–236 (2002)
3. Michiardi, P., Molva, R.: Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In: Proc. 6th Conference on Security Communications, and Multimedia (CMS 2002), pp. 107–121 (2002)
4. Buchegger, S., Boudec, J.Y.L.: The effect of rumor spreading in reputation systems for mobile ad-hoc networks. In: Proc. Workshop on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt 2003), pp. 131–140 (2003)
5. He, Q., Dapeng, W., Khosla, P.: Sori: a secure and objective reputation-based incentive scheme for ad-hoc networks. In: Proc. Wireless Communications and Networking Conference (WCNC 2004), vol. 2, pp. 825–830 (2004)
6. Buchegger, S., Boudec, J.Y.L.: Self-policing mobile ad hoc networks by reputation systems. IEEE Communications Magazine, Special Topic on Advances in Self-Organizing Networks 43(7), 101–107 (2005)
7. Seredynski, M., Bouvry, P.: Evolutionary game theoretical analysis of reputation-based packet forwarding in civilian mobile ad hoc networks. In: Proc. The 22th IEEE International Parallel and Distributed Processing Symposium, NIDISC Workshop (May 2009)
8. Yan, L., Hailes, S.: Cooperative packet relaying model for wireless ad hoc networks. In: Proc. 1st ACM International Workshop on Foundations of Wireless Ad Hoc and Sensor Networking and Computing, pp. 93–100. ACM, New York (2008)
9. Milan, F., Jaramillo, J., Srikant, R.: Achieving cooperation in multihop wireless networks of selfish nodes. In: Proc. Workshop on Game Theory for Communications and Networks. ACM, New York (2006)
10. Yan, L., Hailes, S.: Designing incentive packet relaying strategies for wireless ad hoc networks with game theory. In: Wireless Sensor and Actor Networks II, pp. 137–148. Springer, Boston (2008)
11. Marti, S., Giuli, T., Lai, K., Baker, M.: Mitigating routing misbehavior in mobile ad hoc networks. In: Proc. ACM/IEEE 6th International Conference on Mobile Computing and Networking (MobiCom 2000), pp. 255–265 (2000)

A Comparison of Metaheuristics for the Problem of Solving Parametric Interval Linear Systems

Iwona Skalna and Jerzy Duda

AGH University of Science and Technology, Krakow, Poland
skalna@agh.edu.pl, jduda@zarz.agh.edu.pl

Abstract. The problem of computing a hull solution of parametric interval linear systems with general dependencies is considered. It can be reduced to the problem of solving a family of constrained optimization problems. In this study, different metaheuristics are used to solve those problems. Comparison of evolutionary algorithm, simulated annealing and tabu search algorithm together with analysis of variance tests are provided on the basis of three different practical problems.

1 Introduction

Many real-life problems involve imprecision, approximation, or uncertainty. When the information about uncertain parameter in a form of probability distribution is not available, then the interval analysis can be used most conveniently. In interval analysis, an unknown model parameter $\tilde{p}$ is replaced by an interval number $\mathbf{p}$. Then, the radius $r(\mathbf{p})$ is a measure for the absolute accuracy of the midpoint $\tilde{p}$ considered as an approximation of an unknown value $\tilde{p}$ contained in $\mathbf{p}$.

If a problem is described by a system of linear equations and some of the parameters are unknown but bounded, the problem can be transformed into a parametric interval linear system (PILS). Generally, the solution set of PILS has a very complicated shape [3]. Therefore, the problem of solving PILS is usually reduced to a problem of finding an interval vector, called *outer solution*, that contains the solution set, and the goal is to be as narrow as possible. The tightest outer solution is called *interval hull solution* [7] or *optimal solution* [10]. Preliminaries to a theory are described in Section 2.

The problem of computing the interval hull solution of a parametric interval linear system can be formulated as a problem of solving $2n$ optimization problems, as described in Section 3. They will be solved using metaheuristic strategies such as: evolutionary optimization, simulated annealing, and tabu search, which are presented in Section 4. Metaheuristics are used in hope of obtaining an efficient method for solving parametric interval linear systems with large uncertainties (note, that the problem of computing hull solution is NP-hard in general [9]). The necessity for such methods results from the fact that most of the methods described in the literature produce overestimation of outer solution

and underestimation of inner solution, which increase along with the width of parameter intervals. Solutions obtained using metaheuristics approximates a hull solution from below. It will be shown in Section 5 that those approximations are very close to the actual hull, and hence may be very useful for analysis of a given problem. The main advantage of the metaheuristics is that they can solve PILS with very complicated dependencies and no preliminary transformations are required. The performance of the proposed strategies is verified using several real-life cases.

2 Interval Equations

Consider the interval equation

$$F(u, \mathbf{p}) = 0, \quad (1)$$

where $\mathbf{p}$ is a k -dimensional vector of interval parameters, $u = (u_1, \dots, u_n)$, and $F = (F_1, \dots, F_n)$.

The solution set of system (1) is defined as

$$u^p = \{u : F(u, p) = 0, p \in \mathbf{p}\}. \quad (2)$$

Generally, instead of the solution set u^p itself, various interval solutions are calculated. The interval hull of u^p is called an *interval hull solution* and is defined as $\square u^p = [\inf u^p, \sup u^p]$. Any other interval vector x' such that $x' \supseteq \square u^p$ is called *outer solution*. Similarly, any interval vector x'' such that $x'' \subseteq \square u^p$ is referred to as *inner solution*.

3 Optimization Problems

The problem of computing the hull solution $\square u^p$ can be defined as a problem of solving a family of the following $2n$ global optimization problems:

$$\begin{aligned} \min\{u_i : F(u, p) = 0, p \in \mathbf{p}\} \\ \max\{u_i : F(u, p) = 0, p \in \mathbf{p}\}, \quad i = 1, \dots, n, \end{aligned} \quad (3)$$

and the following theorem holds.

Theorem 1. *Let $F(u, \mathbf{p}) = 0$ and let $\underline{u}_i$ and $\overline{u}_i$ denote, respectively, the solution of the i -th minimization and maximization problem (3). Then*

$$\square u^p = [\underline{u}_1, \overline{u}_1] \times \dots \times [\underline{u}_n, \overline{u}_n]. \quad (4)$$

Global optimization problems (3) can be solved using metaheuristics. Clearly, the result will approximate the hull solution from below, but, as it will be shown, a high quality approximation can be obtained.

4 Metaheuristics

Population P consists of pop_{size} individuals characterized by k -dimensional vectors of parameters $p_i = \{p_{i1}, \dots, p_{ik}\}^T$, where $p_{ij} \in \mathbf{p}_j$, $i = 1, \dots, pop_{size}$, $j = 1, \dots, k$.

4.1 Evolutionary Optimization

Elements of the initial population are generated at random based on the uniform distribution. The 10% of the best individuals pass to the next generation, and the rest of population is generated using the *non-uniform mutation*

$$p'_j = \begin{cases} p_j + (\bar{p}_j - p_j) r (1 - t/n)^b, & \text{if } q < 0.5 \\ p_j + (p_j - \underline{p}_j) r (1 - t/n)^b, & \text{if } q \geq 0.5 \end{cases},$$

and *arithmetic crossover*

$$p^{1'} = rp^1 + (1 - r)p^2, \quad p^{2'} = rp^2 + (1 - r)p^1,$$

where $r, q \in [0, 1]$ are random numbers, and n is a number of generations. It turns out from numerical experiments that mutation rate r_{mut} should be close to 1, and the crossover rate r_{crs} should be less than 0.3. Population size and number of generations depend strongly on the size problem. Here, $pop_{size} = 16$, $n = 60$ occurred to be enough to obtain a very good approximation at a relatively low computation cost. General outline of the algorithm is shown in Fig. 1.

```

Initialize  $P$  of  $pop_{size}$  at random
while ( $i < n$ ) do
  Select  $P'$  from  $P$ 
  Choose parents  $p_1$  and  $p_2$  from  $P'$ 
  if ( $r_{[0,1]} < r_{crs}$ ) then Offspring  $o_1$  and  $o_2 \leftarrow$  Recombine  $p_1$  and  $p_2$ 
  if ( $r_{[0,1]} < r_{mut}$ ) then Mutate  $o_1$  and  $o_2$ 
end while

```

Fig. 1. Evolutionary algorithm

4.2 Simulated Annealing

Since preliminary results given by standard simulated annealing ([2], [4]) algorithm were poor comparing to the other algorithms tested, the authors developed a modified SA algorithm shown in Fig. 2. Algorithm starts from a solution p , which was chosen as the best solution out of pop_{size} solutions instead of starting from a randomly generated solution. Additionally, two iterations (inner $i = 0, \dots, n$ and outer $j = 0, \dots, m$) are used, and current solution p is reset to the best solution found so far p_{best} after each outer iteration. As a perturbation mechanism, a procedure similar to the non-uniform mutation was applied. A perturbed solution p' was obtained by altering a randomly chosen element p_j from the vector representing solution $p = (p_1, \dots, p_j, \dots, p_k)^T$. In order to give the same computation time restriction as for evolutionary algorithm, the following parameters were taken $pop_{size}=500$, $n=100,000$, and $m=100$. Initial temperature t_0 was set to 0.9, and it was degraded by 0.995 after each iteration.

```

Initialize  $P$  of  $pop_{size}$  at random;  $t = t_0$ 
Find best solution  $p$  from  $P$ :  $p_{best} \leftarrow p$ 
while ( $i < n$ ) do
  while ( $j < m$ ) do
     $p' \leftarrow Perturbate(p)$ 
    if ( $f(p') > f(p_{best})$ ) then  $p \leftarrow p'$   $f(p_{best}) = f(p')$ ;  $p_{best} \leftarrow p$ 
    else if ( $r < exp(|f(p') - f(p_{best})|/t)$ ) then  $p \leftarrow p'$ 
     $Decrease(t)$ 
  end while
   $p \leftarrow p_{best}$ 
end while

```

Fig. 2. Simulated annealing

4.3 Tabu Search

Standard tabu search algorithm (see e.g. [2,4]) is used with no aspiration function, however like for the simulated annealing algorithm, it occurred that it is better to start with relatively good solution (taken as the best solution out of pop_{size} randomly generated solutions) than from a single random solution. The neighbour of solution is created by performing the same perturbation as in the simulated annealing algorithm. An outline of the tabu search algorithm is shown in Fig. 3. As for previous metaheuristics, while choosing parameters the aim was to preserve the same computation time, and after some initial computational experiments following parameters were applied: $pop_{size}=200$, $n=300,000$, and $tabu_{size}=30$.

```

Initialize  $P$  of  $pop_{size}$  at random
Find best solution  $p$  from  $P$ :  $p_{best} \leftarrow p$ 
while ( $i < n$ ) do
  Choose best  $p \in Neighbourhood(p_{best})$ 
  repeat
    Choose  $j$  at random
  until ( $j \notin Tabu$ )
   $p' \leftarrow Perturbate(p)$ ;  $Tabu \leftarrow Tabu \cup j$ 
  if ( $\#Tabu > tabu_{size}$ ) then  $Remove(Tabu_1)$ 
  if ( $f(p') > f(p_{best})$ ) then  $p \leftarrow p'$   $f(p_{best}) = f(p')$ 
end while

```

Fig. 3. Tabu search

5 Numerical Experiments

The performance of the methods described in Section 4 is illustrated by numerical solutions of several real-life cases. For each problem, four solutions are presented: hull solution (HS) and the best solution out of 20 runs of evolutionary optimization (EO), simulated annealing (SA), and tabu search (TS).

Example 1 (Two-bay truss)

Two-bay truss (see Fig. 4) elements have the following data: $A = 0.01 \text{ m}^2$, $E = 200 \text{ GPa}$. In this case, three series of computational experiments have been performed with uncertainty levels 1%, 10% and 20% in Modulus of Elasticity (ME) and 40% uncertainty in load. Selected displacements (horizontal $\mathbf{x}_2$ and vertical $\mathbf{y}_2$ displacements of node 2, and horizontal displacement $\mathbf{x}_4$ of node 4) are given in the following tables.

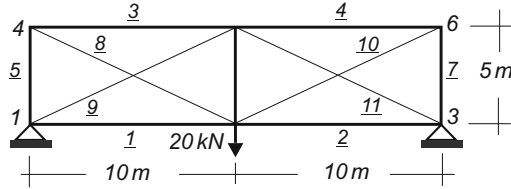


Fig. 4. Two-bay truss

Table 1. Two-bay truss: 1% uncertainty in ME and 40% uncertainty in load

Method	$\mathbf{x}_2 (\times 10^3) [\text{m}]$	$\mathbf{y}_2 (\times 10^3) [\text{m}]$	$\mathbf{x}_4 (\times 10^3) [\text{m}]$
HS	[−0.022556, 0.022556]	[−24.039109, −15.866609]	[3.118214, 4.804920]
EO	[−0.022556, 0.022556]	[−24.039109, −15.866610]	[3.118215, 4.804919]
SA	[−0.021841, 0.022035]	[−24.032914, −15.882790]	[3.118943, 4.796738]
TS	[−0.022470, 0.022550]	[−24.039109, −15.866609]	[3.118214, 4.804919]

Table 2. Two-bay truss: 10% uncertainty in ME and 40% uncertainty in load

Method	$\mathbf{x}_2 (\times 10^3) [\text{m}]$	$\mathbf{y}_2 (\times 10^3) [\text{m}]$	$\mathbf{x}_4 (\times 10^3) [\text{m}]$
HS	[−0.234370, 0.234370]	[−25.177804, −15.186612]	[2.739806, 5.390023]
EO	[−0.234363, 0.234368]	[−25.177796, −15.186614]	[2.739810, 5.390018]
SA	[−0.184566, 0.166457]	[−25.048788, −15.240488]	[2.752346, 5.263116]
TS	[−0.220389, 0.234143]	[−25.177804, −15.186612]	[2.739886, 5.390017]

Table 3. Two-bay truss: 20% uncertainty in ME and 40% uncertainty in load

Method	$\mathbf{x}_2 (\times 10^3) [\text{m}]$	$\mathbf{y}_2 (\times 10^3) [\text{m}]$	$\mathbf{x}_4 (\times 10^3) [\text{m}]$
HS	[−0.490611, 0.490611]	[−26.576571, −14.496311]	[2.323803, 6.067822]
EO	[−0.490603, 0.490599]	[−26.576564, −14.496316]	[2.323825, 6.067812]
SA	[−0.452874, 0.437114]	[−26.492189, −14.656000]	[2.325224, 5.886955]
TS	[−0.479252, 0.484509]	[−26.576570, −14.496312]	[2.323858, 6.065004]

Example 2 (Four-bay truss)

Four-bay truss (see Fig. 5) elements have the same data as in the previous example. Four cases of high uncertainty levels are considered. Bounds for displacements of selected nodes are given in the following tables.

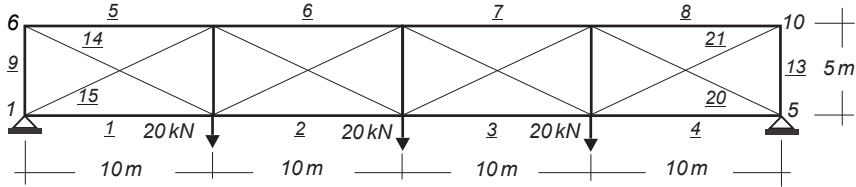


Fig. 5. Four-bay truss

Table 4. Four-bay truss: 10% uncertainty in ME and 40% uncertainty in loads

Method	$\mathbf{x}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{y}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{x}_4 \ (\times 10^2) \ [\text{m}]$
HS	$[-1.323110, -0.665117]$	$[-15.636840, -9.408482]$	$[-0.213123, 0.213123]$
EO	$[-1.322182, -0.665311]$	$[-15.635298, -9.409338]$	$[-0.212346, 0.212550]$
SA	$[-1.297941, -0.688587]$	$[-15.484739, -9.538618]$	$[-0.113371, 0.118707]$
TS	$[-1.318527, -0.693429]$	$[-15.626150, -9.408771]$	$[-0.205510, 0.204041]$

Table 5. Four-bay truss: 20% uncertainty in ME and 40% uncertainty in loads

Method	$\mathbf{x}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{y}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{x}_4 \ (\times 10^2) \ [\text{m}]$
HS	$[-1.508556, -0.567596]$	$[-16.525016, -8.966022]$	$[-0.431053, 0.431053]$
EO	$[-1.507705, -0.568091]$	$[-16.523292, -8.967277]$	$[-0.429775, 0.429693]$
SA	$[-1.434622, -0.604837]$	$[-16.277392, -9.045541]$	$[-0.286158, 0.282184]$
TS	$[-1.505930, -0.579227]$	$[-16.521675, -8.970046]$	$[-0.413044, 0.417879]$

Table 6. Four-bay truss: 40% uncertainty in ME and 60% uncertainty in loads

Method	$\mathbf{x}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{y}_2 \ (\times 10^2) \ [\text{m}]$	$\mathbf{x}_4 \ (\times 10^2) \ [\text{m}]$
HS	$[-2.103750, -0.339409]$	$[-20.185748, -7.157695]$	$[-0.970212, 0.970212]$
EO	$[-2.101508, -0.340497]$	$[-20.175105, -7.159246]$	$[-0.967159, 0.968365]$
SA	$[-1.988016, -0.421076]$	$[-19.933124, -7.264063]$	$[-0.435005, 0.463360]$
TS	$[-1.878610, -0.345227]$	$[-19.980943, -7.180316]$	$[-0.905818, 0.949845]$

Table 7. Four-bay truss: 60% uncertainty in ME and 80% uncertainty in loads

Method	$\mathbf{x}_2 \ (\times 10^2) \text{ [m]}$	$\mathbf{y}_2 \ (\times 10^2) \text{ [m]}$	$\mathbf{x}_4 \ (\times 10^2) \text{ [m]}$
HS	[−2.699244, −0.194814]	[−23.119687, −6.559599]	[−1.546489, 1.546489]
EO	[−2.694410, −0.195005]	[−23.107266, −6.562668]	[−1.544995, 1.543776]
SA	[−2.195022, −0.225934]	[−20.933984, −6.627581]	[−0.730826, 0.777577]
TS	[−2.473199, −0.201909]	[−23.052995, −6.564011]	[−1.517235, 1.535394]

Example 3 ([1], [8])

A simple one-bay structural steel frame, originally considered in [1], is presented in Fig. 6. Initially, the problem is solved with parameter uncertainties which are 5% of the nominal values presented in [8] (Example 5.1). Next, the uncertainty is increased to 30% of the nominal values. Notation for the solution components proposed by Popova [8] is used.

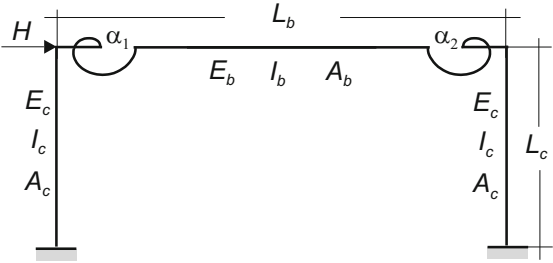


Fig. 6. One-bay structural steel frame

Table 8. One-bay steel frame with $\pm 5\%$ uncertainty in all parameters

Method	$d2_x \text{ [m]}$	$d2_y \ (\times 10^3) \text{ [m]}$
HS	[0.1481510450, 0.1585164153]	[0.3120904116, 0.3419798294]
EO	[0.1481510450, 0.1585164153]	[0.3120904116, 0.3419798294]
SA	[0.1483068563, 0.1584893720]	[0.3136486750, 0.3418576486]
TS	[0.1481691495, 0.1584923325]	[0.3121039574, 0.3418809984]

Table 9. One-bay steel frame with $\pm 30\%$ uncertainty in all parameters

Method	$d2_x \text{ [m]}$	$d2_y \ (\times 10^3) \text{ [m]}$
HS	[0.1243887983, 0.1869551011]	[0.2456606552, 0.4262122935]
EO	[0.1243887983, 0.1869551011]	[0.2456606552, 0.4262122935]
SA	[0.1247561678, 0.1862165051]	[0.2460437763, 0.4228745154]
TS	[0.1243984902, 0.1868282754]	[0.2457210814, 0.4217473869]

6 Conclusions

The paper compares three the most popular metaheuristics applied for solving parametric interval linear system with coefficients that are arbitrary functions of interval parameters. Several numerical experiments showed that Evolutionary Optimization can be quite efficient in solving problems similar to the ones presented in the paper. Also, the tabu search algorithm usually performed quite well giving the results close to those of EO. The simulated annealing, in spite of several modifications, generated usually the worst solutions. Moreover, the TS and SA performed very unstable, and in some runs they received relatively poor solutions. The authors believe further improvements can be made to TS algorithm in order to improve its stability and accuracy.

Acknowledgements. The authors wish to express their sincere thanks to all reviewers for their valuable remarks and suggestions.

References

1. Corliss, G., Foley, C., Kearfott, R.B.: Formulation for Reliable Analysis of Structural Frames. In: Proceedings of NSF Workshop on Reliable Engineering Computing, Savannah, Georgia, USA (2004)
2. Dréo, J., Pétrowski, A., Siarry, P., Taillard, E.: Metaheuristics for Hard Optimization. Springer, Heidelberg (2006)
3. Alefeld, G., Kreinovich, V., Mayer, G.: The Shape of the Solution Set for Systems of Interval Linear Equations with Dependent Coefficients. *Mathematische Nachrichten* 192(1), 23–36 (2006)
4. Luke, S.: Essentials of Metaheuristics. eBook 226 pages (2009), <http://cs.gmu.edu/~sean/book/metaheuristics/>
5. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin (1996)
6. Michalewicz, Z., Fogel, D.: How to Solve It: Modern Heuristics. Springer, Heidelberg (2004)
7. Neumaier, A.: Interval Methods for Systems of Equations. Cambridge University Press, Cambridge (1990)
8. Popova, E.: Solving Linear Systems whose Input Data are Rational Functions of Interval Parameters. In: Boyanov, T., Dimova, S., Georgiev, K., Nikolov, G. (eds.) NMA 2006. LNCS, vol. 4310, pp. 345–352. Springer, Heidelberg (2007)
9. Rohn, J., Kreinovich, V.: Computing exact componentwise bounds on solutions of linear systems with interval data is NP-hard. *SIAM Journal on Matrix Analysis and Applications (SIMAX)* 16, 415–420 (1995)
10. Shary, S.P.: On optimal solution of interval equations. *SIAM Journal on Numerical Analysis* 32(2), 610–630 (1995)

Parametric Approximation of Functions Using Genetic Algorithms: An Example with a Logistic Curve

Fernando Torrecilla-Pinero¹, Jesús A. Torrecilla-Pinero²,
Juan A. Gómez-Pulido¹, Miguel A. Vega-Rodríguez¹,
and Juan M. Sánchez-Pérez¹

¹ Dep. of Technologies of Computers and Communications,
University of Extremadura, Spain

² Dep. of Building, University of Extremadura, Spain

Abstract. Whenever we have a set of discrete measures of a phenomenon and try to find an analytic function which models such phenomenon, we are solving a problem about finding some parameters that minimizes a computable error function. In this way, parameter estimation may be studied as an optimization problem, in which the fitness function we are trying to minimize is the error one. This work try to do that using a genetic algorithm to obtain three parameters of a function. Particularly, we use data about one village population over time to see the goodness of our algorithm.

Keywords: Parameter estimation, functions, genetic algorithms, population, logistic curve.

1 Introduction

In many engineering problems, when you try to estimate data for water supply, the construction of a dam or a road traffic, in example, you should have an estimation over the population that is going to use these elements. As seen in papers like [2] and [3] the parameter estimation can be used too to predict some results in Medicine or floodings.

This way, we should tackle the problem of estimating the population among the years, and at this point we have to talk about the different ways to estimate.

You can not expect very good results with the estimation methods about the future population that are going to be described because many times their goodness is worse when:

- The forecast period increases
- The population of the area decreases
- The population variation speed increases

Next, we will describe a set of methods that are usually used to estimate the population growth (a deeper description can be found in [7]):

1.1 Arithmetic Method

You have to consider the population growth is constant, this is, equivalent to a rect line, with the following equation:

$$P = P_2 + \frac{P_2 - P_1}{t_2 - t_1}(t - t_2) \quad (1)$$

where P_i is the population at time t_i .

1.2 Uniform Growth Percentage

You have to suppose that the growth proportion follows a compound interest law with the following equation:

$$P = P_0 * (1 + K_U)^n \quad (2)$$

where P_0 is the actual population and n is the number of years. You have to be careful with this estimation method because it could give some higher results.

1.3 MOPU (Spain) Proposed Method

It is a particular case of the previous method where K_U is fixed as follows:

- A K_1 average value is calculated with the values of the last decade.
- Similarly K_2 and K_3 are calculated with the values of the last 25 and 50 years respectively.
- It is selected the value K_2 or K_3 , the one that comes closest to K_1 ; this will be denoted as K^+ .
- K_U is fixed with the expression:

$$K_U = \frac{2K_1 + K^+}{3} \quad (3)$$

When $K_U \geq 0.03$ a particular study must be done.

1.4 Geometric Method

It supposes that the community growth is everytime propotional to its population, with the equation:

$$P = P_2 \left(\frac{P_2}{P_1} \right)^{\left(\frac{t - t_2}{t_2 - t_1} \right)} \quad (4)$$

where P_i is the population at time t_i .

As in the previous method, these results should be considered with caution because they are quite optimistic.

1.5 Decreasing Rate of the Growth

The experience says to us that the growth with the previous method does not kept in a long time, but it decreases as the population is closed to the saturation value. The equation is:

$$P_2 - P_1 = (S - P_1)[1 - e^{-K_d(t_2-t_1)}]K + \quad (5)$$

where S is the limit population of the community and K_d is the growth constant.

The problem comes with estimating S and K_d , specially with S when the population is young enough.

It is a very good method for older populations if the parameters have been estimated well.

1.6 Logistic Method or S -Curve

It is based in the fact that, at the beginning the population growth is geometric, then constant and at the last it decreases until it comes to the saturation value, S , with the equation:

$$P = \frac{S}{1 + Me^{bt}} \quad (6)$$

where:

$$\begin{aligned} - S &= \frac{2P_0P_1P_2 - P_1^2(P_0 + P_2)}{P_0P_2 - P_1^2}, \\ - M &= \frac{S - P_0}{P_0}, \\ - b &= \frac{1}{n} \left[\frac{P_0(S - P_1)}{P_1(S - P_0)} \right] \text{ and} \\ - n &= (t_2 - t_1) = (t_1 - t_0). \end{aligned}$$

To calculate S , M and b the populations P_0 , P_1 and P_2 are taken in the equidistant times t_0 , t_1 and t_2 and P_2 is usually taken as the population of the last census.

This method ([9],[10]) is useful to estimate future populations in developed communities, it is usually the most used method and it is the method in which we will put our efforts.

2 Problem Definition

Once explained the methods to estimate the populations, we should focus on studying how to estimate the parameters S , M and b .

The method explains how to calculate this values taking the populations of the three last measures in three moments of the time equidistantly distributed. When you try to adjust the curve with more parameters than just these three values, or taking some times not equidistantly distributed, to make a more realist estimation, you realized it is not possible with this method, so you have to resort

to another way. At this point, we could think about the genetic algorithm as a powerful tool to estimate S , M and b , assuming that the problem could be tackled as an optimization one ([1],[4]).

The whole approach to the problem has been done from the point of view of a real encoding.

2.1 Functions

The main functions in this problem will be:

- Curve: defines the logistic curve like $P = \frac{S}{1+Me^{bt}}$
- Error: defines the average cuadratic error between the estimation and the real value of the parameters.
- Variance: auxiliary function to calculate the variances which will be used in the mutation function.
- Elithism: selects the n best individuals and keep them into the next generation.
- GenerateProb: generates the probability list of each one of the individuals based on their costs.
- Those functions of the genetic algorithms like mutation, combination, selection, cross-over, etc.
 - Combination: combines 2 individuals in terms of their errors.
 - Selection: selects the individual which is below the individual who happens randomly according to the method of roulette.
 - Mutation: mutates one individual in the range $[-var, +var]$ from each variable.

2.2 Fitness-Error Function

The fitness-error function is defined as follows:

$$error = \frac{\sqrt{\sum (ind_i - real_i)^2}}{numyears} \quad (7)$$

being $real_i$ the obtained value when passing the values S , M and b to the curve function for the year i .

We minimize this error to estimate the parameters S , M and b .

3 Experimental Results

To probe the functionality of this work, we have studied some cases taking in example the population growth in towns and villages of the province of Cáceres (Spain). Here are the results for the more representative cases:

As you can see in the fig. 1, the parameter estimation provides a logistic curve that fits better the closer the population data are.

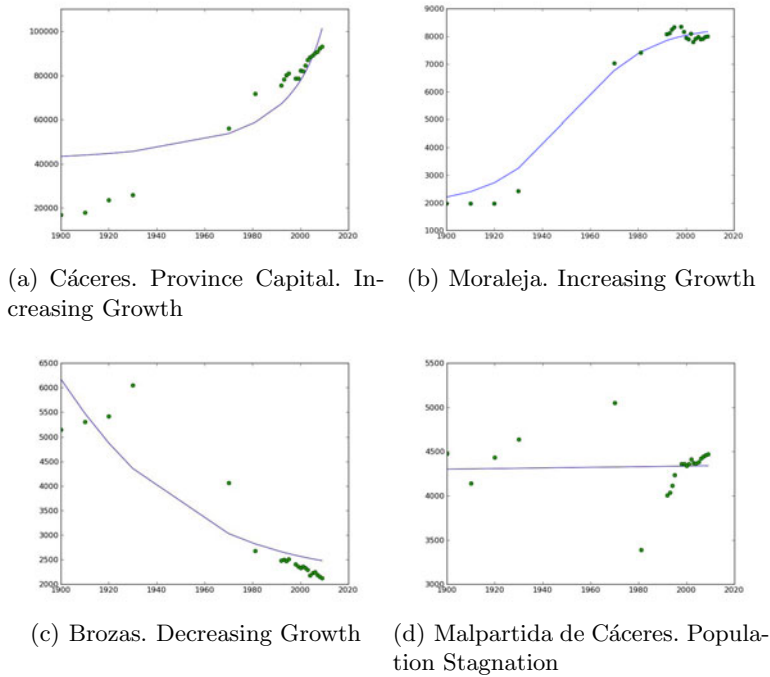


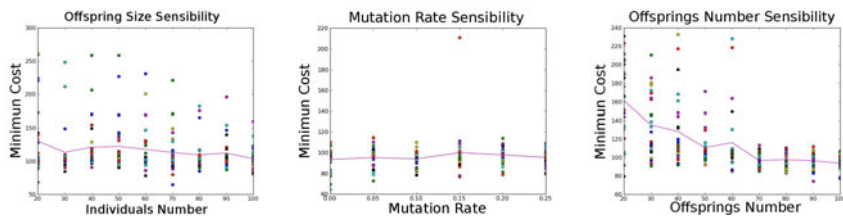
Fig. 1. Logistic Curve in the Province of Cáceres

3.1 Sensibility Analysis

We have made a sensibility analysis involving three main points:

- Sensibility to the offspring size

On the fig. 2(a), you can see that the bigger number of individuals, the smaller median cost is, but the cost reduction is not very significant.



- (a) Sensibility to the offspring size
- (b) Sensibility to the mutation rate
- (c) Sensibility to the number of generations

Fig. 2. Sensibility Analysis

Table 1. Offspring Size Sensibility

Offsprings	50								
Individuals Number	20	30	40	50	60	70	80	90	100
Mutation rate	0.1								
Average Cost	130.21	113.56	120.90	122.36	117.89	113.14	109.56	112.39	103.77

- Sensibility to the mutation rate
When thinking about the importance of the mutation rate (fig. 2(b)), we can see that this parameter can be practically neglected, because it hardly affects the obtained results.
- Sensibility to the offsprings number
Regarding the number of offsprings (fig. 2(c)), it is easy to see that the bigger number of offsprings, the minimum cost decreases. It can be appreciated that from twenty offsprings to seventy, the decrease is significant but beyond it is not very significant. We can set this value as a good value for the parameter when executing the algorithm.

Table 2. Mutation Rate Sensibility

Offsprings	100					
Individuals Number	100					
Mutation Rate	0.00	0.05	0.10	0.15	0.20	0.25
Average Cost	93.41	95.21	94.04	99.94	97.85	95.40

Table 3. Offsprings Number Sensibility

Offsprings Number	20	30	40	50	60	70	80	90	100
Individuals Number	100								
Mutation rate	0.1								
Average Cost	160.90	135.36	128.25	110.77	116.19	96.91	97.73	96.53	93.84

3.2 Comparing Methods and Results

In the fig. 3 you can see a comparison between the MOPU and the GA Logistic Curve we have obtained. In the X axis it is represented the obtained difference with the GA Logistic Curve referring to the real measure of the population; in the Y axis it is represented the obtained difference with MOPU method referring to the real measure of the population.

As you can see, all the points are at or above the bisector, which means that every measures that are made with the GA Logistic Curve are, at least, as good as the measures obtained with the MOPU method.

Comparing with other results, like the one seen in [11], where the reached improvement are between 0 and 2%, we have reached an improvement between 0 and 8.5%.

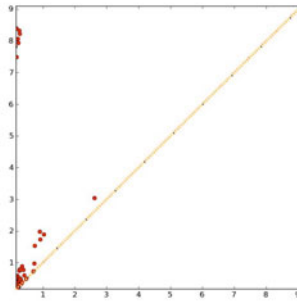


Fig. 3. Comparison between MOPU and GA Logistic Curve

4 Conclusions

Heuristic techniques are not methods restricted in application scope to the optimization, but they have some other applications, and can be applied to different engineering or science fields. In some cases, the genetic algorithms represent an advantageous alternative in terms of necessary computing power and obtained results compared with more traditional techniques, and sometimes could give better results in terms of parameter estimation.

The most important parameters when looking at the achieved improvement in the optimization are the individuals number and, overall, the mutation rate. The efficiency variation with respect to the iterations number with the improved operators is irrelevant, since these operators make a very high pressure that makes the most important improvements to be produced at the first iterations.

Acknowledgment

This work was partially funded by the Spanish Ministry of Science and Innovation and ERDF (the European Regional Development Fund), under the contract TIN2008-06491-C04-04 (the MSTAR project).

References

1. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
2. Jaffrezic, F., Meza, C., Lavielle, M., Foulley, J.-L.: Genetic analysis of growth curves using the SAEM algorithm. *Genet. Sel. Evol.* 38(EDP Sciences), 583–600 (2006)
3. Liu, X.-Y.: An improvement logistic model based on multiple objective genetic algorithm. In: *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, Baoding, July 12–15, pp. 2292–2295 (2009)
4. Syswerda, G.: Schedule optimization using genetic algorithms. In: Davis, L. (ed.) *Handbook of Genetic Algorithms*, pp. 332–349. Van Nostrand Reinhold, New York (1991)

5. Whitley, D.: A genetic algorithm tutorial (2005)
6. Kuhn, E., Lavielle, M.: Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.* 49, 1020–1038 (2005)
7. Universidad Nacional de Colombia: Estimación de la Población Futura, http://www.virtual.unal.edu.co/cursos/sedes/manizales/4080004/contenido/Capitulo_4/Pages/caudales_continuacion1.htm
8. Veres Ferrer, E.: Nuevo procedimiento para el ajuste de la curva logística: aplicación a la población española. *Estadística Española* 108, 5–17 (1985)
9. Martínez, E.: Dinámica poblacional (II): la ecuación logística, <http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/calculo/poblacion/logistica/logistica.html>
10. Poveda, R.G., Manrique, H.J.: Aplicación de la curva logística a los censos de la ciudad de Medellín. *Ecos de Economía* (25) (2007)
11. Vinterbo, S., Ohno-Machado, L.: A Genetic Algorithm to Select Variables in Logistic Regression: Example in the Domain of Myocardial Infarction. In: AMIA, Inc., pp. 984–988 (1999)

Population-Based Metaheuristics for Tasks Scheduling in Heterogeneous Distributed Systems

Flavia Zamfirache, Marc Frîncu, and Daniela Zaharie

Department of Computer Science, West University of Timișoara, Romania
{zflavia,mfrincu,dzaharie}@info.uvt.ro

Abstract. This paper proposes a simple population based heuristic for task scheduling in heterogeneous distributed systems. The heuristic is based on a hybrid perturbation operator which combines greedy and random strategies in order to ensure local improvement of the schedules. The behaviour of the scheduling algorithm is tested for batch and online scheduling problems and is compared with other scheduling heuristics.

1 Introduction

Since the work of Braun et al. [1] which illustrated the fact that genetic algorithms can generate good solutions for task scheduling problems, a lot of other population-based metaheuristics were proposed (e.g. evolutionary algorithms [2], ant systems [7], memetic algorithms [10]). Unlike the genetic algorithm in [1] which is based on classical mutation and crossover operators, the recent approaches use specific local search operators. Most researchers identified as effective operators those involving a rebalancing of the load on different processors by moving or swapping tasks between processors. Currently there exist both simple and sophisticated "rebalancing" operators. The aim of this paper is to identify the basic components of such operators and to design a simple population-based scheduler involving as few as possible search mechanisms.

The addressed problem is that of assigning a set of independent and non-preemptive tasks to a set of resources (e.g. machines, processors) such that the maximal execution time over all resources, i.e. *makespan*, is minimized. The assignment of tasks is based on estimations of the execution times of the tasks on various resources. Let us consider a set of n tasks, $\{t_1, \dots, t_n\}$, to be scheduled on a set of $m < n$ processors, $\{p_1, \dots, p_m\}$. Let us suppose that for each pair (t_i, p_j) we know an estimation $ET(i, j)$ of the time needed to execute the task t_i on the processor p_j . A schedule is an assignment of tasks to resources, $S = (p_{j_1}, \dots, p_{j_n})$, where $j_i \in \{1, \dots, m\}$ and p_{j_i} denotes the processor to which the task t_i is assigned. If $\mathcal{T}_j$ denotes the set of tasks assigned to processor p_j and T_j^0 denotes the time moment since the processor j is free then the completion time corresponding to this processor will be $CT_j = T_j^0 + \sum_{i \in \mathcal{T}_j} ET(i, j)$. The makespan is just the maximal completion time over all processors, i.e. $makespan = \max_{j=\overline{1,m}} CT_j$. The problem to be solved is that of finding the

Table 1. Characteristics of the strategies used to construct initial schedules (Notations: CT - current completion time, ET - execution time, ECT - estimated completion time)

Task selection	Processor selection	Strategy
Random	Random	Random
Random	Min CT	Opportunistic Load Balancing (OLB)
Random	Min ET	Minimum Execution Time (MET)
Random	Min ECT	Minimum Completion Time (MCT)
Increasing min ECT	Min ECT	MinMin
Decreasing min ECT	Min ECT	MaxMin

schedule with the minimal makespan value. In real distributed systems tasks arrive continuously and have to be assigned to resources either as they arrive or when a scheduling event is triggered. In this work we analyze both the case when the scheduling event is triggered when a given number of tasks have arrived (batch scheduling) and the case when the scheduling is activated at pre-specified moments of time (online scheduling). The main idea of the proposed population-based heuristic is described in Section 2. Sections 3 and 4 present the numerical results obtained for batch and online scheduling problems while Section 5 concludes the paper.

2 Designing Heuristics for Task Scheduling

The construction of a (sub)-optimal schedule is usually based on creating an initial schedule which is then iteratively improved. When constructing an initial schedule there are two decisions to take: (i) the order in which the tasks are assigned to processors; (ii) the criterion used to select the processor corresponding to each task. Depending on these elements there exist several strategies [1] as presented in Table 1. Each of these strategies generates initial schedules with a specific potential of being improved. Therefore it would be beneficial to use not just one strategy but to use a population of initial schedules constructed through different strategies.

The initial schedules created by the scheduling heuristics are usually non-optimal and thus they can be improved by moving or swapping tasks between resources. Depending on the criteria used to select the source and destination resources, and the tasks to be relocated there can be designed a lot of strategies to perturb a schedule [9]. Most perturbation operators involved in the scheduling heuristics used in task scheduling are based on two typical operations: "move" one task from a resource to another one and "swap" two tasks between their resources. In order to obtain an immediate improvement in the schedule, the most loaded resource (which determine the makespan) should be involved in the operation. The largest improvement can be obtained by an exhaustive search for the pair consisting of the task to be moved and the destination processor. Besides the fact that this operation is costly ($\mathcal{O}(mn)$) it can fail to generate in just one step a schedule with a smaller makespan. For instance, if there are several

Table 2. Characteristics of the strategies used to perturb the schedules

Source		Destination		Strategy
Processor	Task	Processor	Task	
Random	Random	Random	-	<i>Random Move</i>
Most loaded (max CT)	Random	Best improvement	-	<i>Greedy Move</i>
Most loaded (max CT)	Random	Least Loaded (min CT)	Random	<i>Greedy Swap</i>

processors reaching the maximal completion time it is necessary to apply for several times the "move" operation in order to obtain a decrease of the overall makespan. On the other hand if there is no pair (task, destination processor) which allows to decrease the makespan of the source processor then the "swap" operation should be used instead. In the case of an exhaustive search for the pair of tasks to be swapped the complexity order could be in the worst case $\mathcal{O}(n^2)$ which for a large number of tasks becomes impractical. Therefore from the large number of possible choices of source and destination processors and of tasks to be relocated we selected those which do not involve a systematic search in the set of tasks (i.e. the tasks to be relocated are randomly chosen). The strategies presented in Table 2 were selected based on their simplicity, efficiency and randomness/greediness balance. The "random move" corresponds to the "local move" operator [10] and is similar to the mutation operator used in evolutionary algorithms. The "greedy move" operator is related to the "steepest local move" in [10] but with a higher greediness since it always involves the most loaded processor. The "greedy swap" is similar to "steepest local swap" in [10] but it is less greedy and less expensive since it does not involve a search over the set of tasks.

Since one perturbation step does not necessarily lead to an improvement in the quality of a schedule we consider an iterated application of the perturbation step until either n iterations were executed (each task has the chance to be moved) or a maximal number, g_p , of unsuccessful perturbations is reached.

The influence of g_p on the quality of the schedule is analyzed in the next section. On the other hand in order to exploit the search abilities of each strategy it seems natural to combine several perturbation operators. Thus the strategies in Table 2 are combined as described in the Algorithm 1 (*HybridPerturbation*). This hybrid perturbation has a structure similar to the "re-balancing" mutation described in [10]. However there are some differences between them. In [10] the "swap" perturbation is applied before "move" perturbation while in the hybrid perturbation described in Algorithm 1 the order is reversed. This apparently minor difference influences the overall cost of the perturbation as the application of the "move" operation is less costly than that of "swap" and it can induce a larger gain in the makespan. On the other hand in [10] only one perturbation step is applied to a schedule at each evolutionary generation. Moreover in the "re-balancing" operator the random perturbation is applied any time when the "swap"- "move" duo is unsuccessful while in our case the random perturbation is interpreted as a mutation, thus it is applied with a small probability (e.g. $p_m = 1/n$).

Algorithm 1. The general structure of the population based scheduler*SimplePopulationScheduler*
(SPS)

```

1: Generate the set of initial schedules:
2:  $S \leftarrow \{S_1, \dots, S_N\}$ 
3: while (the stopping condition is false) do
4:   for  $i = 1, \overline{N}$  do
5:      $S'_i \leftarrow \text{perturb}(S_i)$ 
6:   end for
7:    $S \leftarrow \text{select}(S, \{S'_1, \dots, S'_N\})$ 
8: end while

```

SimplePerturbation(S)

```

1:  $i \leftarrow 0$ ; fail  $\leftarrow 0$ 
2: while  $i < n$  and fail  $< g_p$  do
3:    $i \leftarrow i + 1$ 
4:   if GreedyMove/Swap( $S$ ) is successfull then
5:     fail  $\leftarrow 0$ ;  $S \leftarrow \text{GreedyMove/Swap}(S)$ 
6:   else
7:     fail  $\leftarrow \text{fail} + 1$ 
8:     if random(0, 1)  $< p_m$  then
9:        $S \leftarrow \text{RandomMove}(S)$ 
10:    end if
11:  end if
12: end while
13: return  $S$ 

```

HybridPerturbation(S)

```

1:  $i \leftarrow 0$ ; fail  $\leftarrow 0$ 
2: while  $i < n$  and fail  $< g_p$  do
3:    $i \leftarrow i + 1$ 
4:   if GreedyMove( $S$ ) is successfull then
5:     fail  $\leftarrow 0$ ;  $S \leftarrow \text{GreedyMove}(S)$ 
6:   else
7:     if GreedySwap( $S$ ) is successfull then
8:       fail  $\leftarrow 0$ ;  $S \leftarrow \text{GreedySwap}(S)$ 
9:     else
10:      fail  $\leftarrow \text{fail} + 1$ 
11:      if random(0, 1)  $< p_m$  then
12:         $S \leftarrow \text{RandomMove}(S)$ 
13:      end if
14:    end if
15:  end if
16: end while
17: return  $S$ 

```

Having the perturbation as key operator we designed a simple population-based heuristics described in Algorithm 1 (*SPS* - *SimplePopulationScheduler*). Besides the perturbation operator which can be a simple (*SimplePerturbation*) or a hybrid one (*HybridPerturbation*) there are two other elements which can influence the behaviour of the algorithm: initialization and selection. The use of some seed schedules in the initial population has been emphasized by many authors [1,6,10]. Consequently, besides the plain random schedules we included in the initial population also schedules generated with the heuristics listed in Table 1. During the iterative process, each schedule, S_i , in the current population is perturbed leading to a new schedule S'_i (it should be mentioned that in the case of unsuccessful perturbation, S_i could remain unchanged). The schedules corresponding to the next iterative step (generation) are selected from the sets of current and perturbed schedules using a binary tournament approach (the schedule with the smallest makespan from a randomly selected pair of schedules is selected). To ensure the elitism, the best element of the population is preserved. A preliminary analysis on the role of crossover in generating good schedules illustrated that no significant gain is obtained by using crossover (at least uniform and one cut-point crossover). Since the number of processors is usually significantly smaller than the number of tasks almost all processors are involved in the

schedules included in the population. Thus the set of schedules generated by a crossover operator would not be significantly different from the set of schedules which could be generated by applying only the iterated perturbation.

3 Numerical Results for Batch Scheduling

Let us consider the case where the scheduling event is activated when a given number of tasks arrived to the scheduler. This is a classical batch scheduling problem characterized by the fact that some data concerning the estimated execution time of tasks on different resources is known. As test data we have used those introduced in [1] which provides matrices containing values of the expected computation time (ET) generated based on different assumptions related to tasks and resources heterogeneity (low and high) and consistency (consistent, semi-consistent and inconsistent). The data correspond to the case of 512 tasks to be scheduled on 16 processors.

The aim of the numerical study was to analyse the influence of the perturbation strategies on the performance of a Simple Population-based Scheduler (SPS) having the structure described in Algorithm 1. The parameters involved in the algorithm were set based on preliminary parameters tuning leading to the following values: (i) 25 elements in the population (populations of sizes 50, 100 and 200 were also analysed); (ii) a maximal number of successive failures (g_p) in the perturbation operator equal to 150 (values between 10 and 300 were tested; the influence of this parameter on the performance of the scheduler is illustrated in Figure 1 for three test cases); (iii) a probability of applying random perturbations (p_m) equal to $1/n \simeq 0.002$. The maximal number of iterations involved in the stopping conditions was set to 8000. This is in accordance with the values used in literature for evolutionary schedulers [1]. The average time needed to generate a schedule is around 40s (on a Intel P8400 at 2.26GHz) which is also consistent with the time reported in [10] (90s on a AMD K6(tm) at 450MHz).

The analysed initialization strategies are: (i) random initialization; (ii) use of the scheduling heuristics described in Table 1 and randomly initialize the other elements; (iii) use random perturbations of the scheduling heuristics in Table 1; (iv) use the MinMin heuristic and random perturbations of this. As expected, the best results were obtained when the initial population contains seeds obtained by using scheduling heuristics while the worst behaviour corresponds to purely random initialization. The numerical results presented in Table 3 correspond to the three perturbation variants (move-based, swap-based and the hybrid one) and to a state of the art memetic algorithm hybridized with Tabu Search (MA+TS) [10]. Even if based on simpler operators, the algorithm proposed in this work provides schedules close in quality to those generated by MA+TS. Moreover in the case of inconsistent test cases ("u.i.*") problems) the proposed scheduler using the hybrid perturbation operator provides better results.

Table 3. Averages and standard deviations (computed by 30 independent runs) of the makespan obtained by the population-based scheduler with different perturbation strategies. The best and the second best values (validated by a t-test with a significance level of 0.05) for each problem are in bold and in italic, respectively.

Pb.	GreedyMove	GreedySwap	Hybrid	MA+TS[10]
u_c_hihi	7684852.40(± 24798)	7689131.76(± 26971)	<i>7609663.13</i> (± 30673)	7530020.18
u_c_hilo	155248.33(± 551)	155495.10(± 158)	<i>154979.43</i> (± 180)	153917.17
u_c_lohi	251445.60(± 809)	250558.63(± 1028)	<i>248903.70</i> (± 1014)	245288.94
u_c_lolo	5255.06(± 8.90)	5258.4(± 7.65)	<i>5235.00</i> (± 5.32)	5173.72
u_i_hihi	3072453.70(± 18667)	3019756 (± 14323)	3014083.63 (± 21420)	<i>3058474.90</i>
u_i_hilo	75222.90(± 318.46)	<i>74684.433</i> (± 225.12)	74553.20 (± 130.78)	75108.49
u_i_lohi	106309.56(± 706.70)	<i>105261.20</i> (± 561.41)	105013.60 (± 516.63)	105808.58
u_i_lolo	2617.26(± 16.04)	<i>2590.83</i> (± 8.18)	2585.70 (± 6.05)	2596.57
u_s_hihi	4382845.80(± 50248)	4352017.96(± 36899)	4316556.23 (± 29236)	4321015.44
u_s_hilo	<i>98036.16</i> (± 241.10)	98302.366(± 363.54)	<i>97964.86</i> (± 364.56)	97177.29
u_s_lohi	127565.00(± 613.97)	<i>127026.63</i> (± 499.45)	126763.23 (± 564.75)	127633.02
u_s_lolo	3538.96(± 19.28)	3526.53(± 11.24)	<i>3520.80</i> (± 11.39)	3484.08

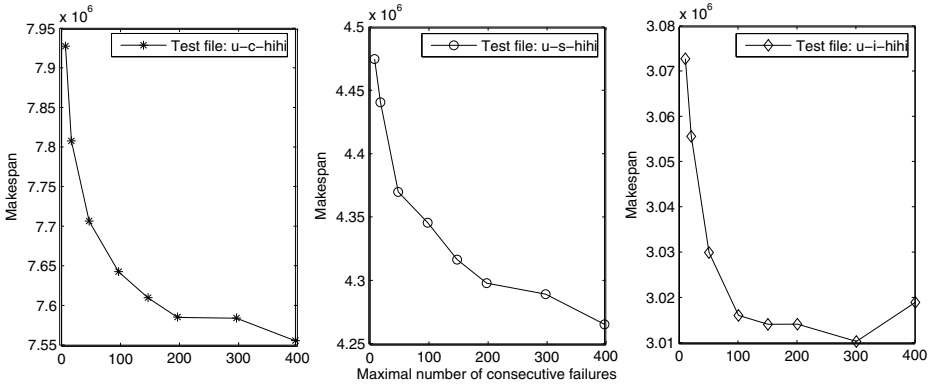


Fig. 1. Influence of the maximal number of consecutive mutations without improvement (g_p) on the makespan

4 Numerical Results for Online Scheduling

For online scheduling we considered a simulation model where task executions times (ET) follow a Pareto distribution with $\alpha = 2$ and the tasks arrival rate is modelled based on statistical results extrapolated from real world traces [3]. A total number of 500 tasks were generated for every test. Rescheduling was done every 250 time units given a minimal execution time of 1000 units. All tests were repeated 20 times in order to collect statistics. The main aim of the numerical tests was to analyze if using populations of schedules one can obtain improvements in the quality with an acceptable loss in the scheduling time.

Therefore several dynamic scheduling heuristics with ageing have been tested against their corresponding population based versions which were constructed by using the specific scheduling heuristics as perturbation operators in SPS. Their behaviour has also been compared with the SPS algorithm based on a non-iterated hybrid perturbation (at each perturbation step the hybrid perturbation is applied only once). Among the online scheduling algorithms we tested a flavour of DMECT as described in [4], the MinQL heuristic [5] and the classic MinMin and MaxMin with ageing. DMECT periodically computes for every task the Local Waiting Time (LWT) - the time since it was assigned to the current processor queue - and a σ value that depends on the implementation and could take into account the estimated execution time (ET). This paper uses the values given in [4]. From these values a decision on whether to move the task or not is taken by checking if the $\sigma - LWT$ is smaller than 0 or not. MinQL allows for optimal balancing the tasks inside resources while taking into account both the age of the task and the priority of local tasks. It uses a backfilling approach where multiple selection conditions for the destination resource can be used. The version used for testing in this paper uses a selection based on the CPU speed.

The population variants of the two previously mentioned scheduling heuristics use a population of 25 elements initialized both with random schedules (60%) and by using the MinMin heuristics (40%). The scheduling heuristics is then applied on every element to generate perturbed schedules and the surviving elements are selected by tournament. The procedure stops when an improvement in the makespan of at least 10% is no longer noticed after a given number of iterations (e.g. 600). Table 4 presents the main benefits of population based scheduling heuristics (pDMECT and pMinQL) when used in online scheduling. Both pDMECT and pMinQL obtained significantly better results than their non-population variants, with pDMECT having a behaviour similar to SPS (the best values in Table 4 are bold-faced and they were validated using a t-test with 0.05 as level of significance). The only notable difference in the behaviour of pDMECT and SPS was that of speed. pDMECT required almost 30 seconds to build a schedule while the simple population-based scheduler needed only three seconds on average. The reason for this difference lies in the complexity of one scheduling step in pDMECT, $\mathcal{O}(m \times n)$, compared with that of one perturbation step in SPS, $\mathcal{O}(m)$, where m represents the number of processors and is significantly smaller than n which is the number of tasks.

Table 4. Average makespan (MS) obtained by online scheduling heuristics and their population based variants

	DMECT	pDMECT	MinQL	pMinQL	SPS	MaxMin	MinMin
MS	66556.20± 15097.85	49409.11± 9522.13	76564.40± 18114.51	54332.89± 9891.15	46996.76± 8812.87	61165.15± 11936.19	68774.87± 15101.05
Time (ms)	66.56 ± 15.50	28343.04± 10702.15	3.06 ± 2.52	2254.64± 314.45	2777.70± 578.22	684.49± 242.15	669.21± 209.99

5 Conclusions

The simple population-based scheduler using an iterated hybrid perturbation operator provides solutions to batch scheduling problems which are comparable in quality with those generated by schedulers using more sophisticated local search operators [10]. The main benefit is obtained in the case of highly heterogeneous and inconsistent distributed environments. The idea of using a simple population-based heuristic proved to ensure a good compromise between solution quality and computational cost also in the case of online scheduling. Further work will address the case of interrelated tasks and that of using other metrics such as the *Total Processing Consumption Cycle* which is an alternative to the makespan and is independent of the hardware.

Acknowledgments. This work is supported by Romanian project PNCD II 11-028/14.09.2007 (NatComp).

References

1. Braun, T.D., Siegel, H.J., Beck, N., et al.: A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. *Journal of Parallel and Distributed Computing* 61(6), 810–837 (2001)
2. Carretero, J., Xhafa, F.: Using Genetic Algorithms for Scheduling Jobs in Large Scale Grid Applications. *Journal of Technological and Economic Development - A Research Journal of Vilnius Gediminas Technical University* 12(1), 11–17 (2006)
3. Feitelson, D.G.: Workload modeling for computer systems performance evaluation (2010), <http://www.cs.huji.ac.il/~feit/wlmod/>
4. Frîncu, M.: Dynamic Scheduling Algorithm for Heterogeneous Environments with Regular Task Input from Multiple Requests. In: Abdennadher, N., Petcu, D. (eds.) *GPC 2009*. LNCS, vol. 5529, pp. 199–210. Springer, Heidelberg (2009)
5. Frîncu, M., Macariu, G., Carstea, A.: Dynamic and Adaptive Workflow Execution Platform for Symbolic Computations. *Pollack Periodica, Akademiai Kiado* 4(1), 145–156 (2009)
6. Page, A.J., Keane, T.M., Naughton, T.J.: Multi-heuristic dynamic task allocation using genetic algorithms in a heterogeneous distributed system. *J. Parallel Distrib. Comput.* (2010), doi:10.1016/j.jpdc.2010.03.11
7. Ritchie, G., Levine, J.: A hybrid ant algorithm for scheduling independent jobs in heterogeneous computing environments. In: *Proc. of 23rd Workshop of the UK Planning and Scheduling Special Interest Group* (2004)
8. Page, A.J., Naughton, T.J.: Dynamic task scheduling using genetic algorithms for heterogeneous distributed computing. In: *Proc. of 19th IEEE/ACM International Parallel and Distributed Processing Symposium*, Denver, pp. 1530–2075 (2005)
9. Xhafa, F., Abraham, A.: Computational models and heuristic methods for Grid scheduling problems. *Future Generation Computer Systems* 26, 608–621 (2010)
10. Xhafa, F.: A Hybrid Evolutionary Heuristic for Job Scheduling on Computational Grids. In: *Hybrid Evolutionary Algorithms. Studies in Computational Intelligence*, vol. 75, pp. 269–311. Springer, Heidelberg (2007)

Modeling of Species and Charge Transport in Li-Ion Batteries Based on Non-equilibrium Thermodynamics

Arnulf Latz, Jochen Zausch, and Oleg Iliev

Fraunhofer Institut für Techno- und Wirtschaftsmathematik
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany

Abstract. In order to improve the design of Li ion batteries the complex interplay of various physical phenomena in the active particles of the electrodes and in the electrolyte has to be balanced. The separate transport phenomena in the electrolyte and in the active particle as well as their coupling due to the electrochemical reactions at the interfaces between the electrode particles and the electrolyte will influence the performance and the lifetime of a battery. Any modeling of the complex phenomena during the usage of a battery has therefore to be based on sound physical and chemical principles in order to allow for reliable predictions for the response of the battery to changing load conditions. We will present a modeling approach for the transport processes in the electrolyte and the electrodes based on non-equilibrium thermodynamics and transport theory. The assumption of local charge neutrality, which is known to be valid in concentrated electrolytes, is explicitly used to identify the independent thermodynamic variables and fluxes. The theory guarantees strictly positive entropy production. Differences to other theories will be discussed.

1 Introduction

Mathematical modeling of Li-ion batteries on cell level was pioneered by the work of Newman and his coworkers [1,2,3] and extended and refined by many other authors [4,5,6]. The modeling approach is based on transport equations for Li ions and charges in the electrolyte as well as in the active particles of cathode and anode (for an illustration of the Lithium Ion battery see Fig. 1). Originally the electrodes were considered as porous media [1] made of a porous active particle skeleton filled with electrolyte. Later the porous model was derived with the help of volume averaging techniques for some set of equations for the different transport mechanisms in electrolyte and in the solid active particles [7]. The transport of charges and species between the electrolyte and the active particles was described with the help of a Butler - Volmer reaction model [2] and some assumptions about continuity conditions for charge and species flux. So far approaches where the active particles are resolved and the transport in particles and electrolytes are treated separately are rare [8,9]. But whether one starts directly with the porous electrode model or with a model resolving the

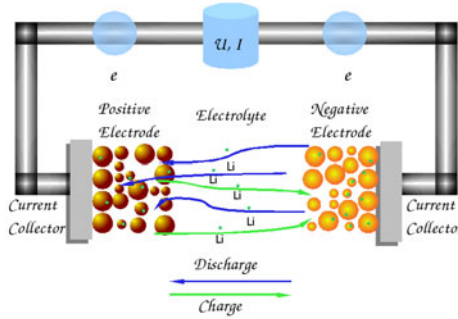


Fig. 1. Illustration of a Lithium ion battery. The spheres are the active particles of the porous electrodes in which Lithium Ions (green dots) can be stored. The voids between particles and between the electrodes are filled with an electrolyte in which Lithium ions diffuse and migrate. Electrons, on the other hand, have to move through the external circuit.

fine structure of the electrode, it is in both cases important to base the battery model on thermodynamically consistent concepts, especially if at some point also the local heat production is to be simulated.

Charge transport mechanisms in active particles and in electrolyte are completely different on the microscopic (atomic) scale. In active particles, charge is transported mainly by pure electronic conduction. The contribution of the ion diffusion in the active particles to the electric current can be neglected due to the large mobility of the electrons compared to the ions. Charge transport in the electrolyte, on the other side, is exclusively due to ionic transport. In fact, the transfer of electrons into the electrolyte would result in the reduction of Li ions in the electrolyte to metallic Lithium, and is considered to be one of the many degradation mechanisms in Li ion batteries [10]. Due to the large mobility of electrons local charge neutrality is easily maintained in the active particles. The charge of an inserted Li Ions is instantaneously shielded by local rearrangements of electronic charges and the transport of electrons into the active particles over the current collectors. In the electrolyte the transport of species (ions) and charge is strongly coupled. Both, charge and species fluxes are caused by gradients in the chemical potential as well as by gradients in the electrical potentials. The constitutive relations for the charge and species fluxes describing these relations are well known for dilute electrolytes [2]. In batteries we have to deal with highly concentrated electrolytes. Some relations for these electrolytes are also derived in [2] combining multicomponent diffusion theory and considerations for chemical equilibria between reacting species in the electrolyte. As it will be shown below, the result is at variance with the general form of constitutive relations in ionic liquids, usually obtained in non equilibrium thermodynamics [11,12]. A main contribution of our paper is therefore the careful rederivation of the constitutive relations for ion and charge flux in a mixture of a fully dissociated binary salt in a neutral solvent, using the well known concepts

of nonequilibrium thermodynamics [11,12]. As it is known that local charge neutrality is preserved in concentrated electrolytes except for the diffuse part of the double layer around active particles [2], we make explicit use of this property in our derivation of the constitutive relations. Due to the limited space for this article, details of the more general derivation, including thermal fluctuations will be given elsewhere [13]. With these equations the transition to the effective porous medium theory for cathode and anode can be obtained with standard techniques as e.g. volume averaging [7].

2 Model

The starting point for a continuum model of charge and species transport in a Li-ion battery are the conservation equation for Li-ion concentration c and charges q . The continuity equation for the concentration of Li Ions c is given by

$$\frac{\partial c}{\partial t} = -\nabla \cdot \mathbf{N}_+ \quad (1)$$

Here $\mathbf{N}_+$ is the flux of Li ions. The equation for the charge concentration is given by

$$\frac{\partial q}{\partial t} = -\nabla \cdot \mathbf{j} \quad (2)$$

where $\mathbf{j}$ is the electrical current. The approximation of charge neutrality requires not only that the time derivative in (2) is identical zero, but that the local charge q vanishes i.e. $q \equiv 0$. The main challenge for a constitutive theory is to derive a thermodynamically consistent relation for the fluxes $\mathbf{N}_+$ and the electrical current $\mathbf{j}$. Also, the influence of solvent molecules and negative ions to the transport properties has to be clarified.

2.1 Charge and Species Transport in a Concentrated Electrolyte

To obtain a thermodynamically consistent model for charge and ion fluxes in the electrolyte we apply the well known formalism of non-equilibrium thermodynamics [11,12] to a mixture of fully dissociated binary salt and a solvent. The concentrations of positive and negative ions with charge z_+ and z_- are c_+ and c_- , respectively. The concentration of the solvent is c_0 . Instead of motivating our theory with considerations from dilute electrolyte theory, we are considering the opposite limit of concentrated electrolytes. In this limit the Debye length λ_D is so small, that local charge fluctuations are restricted to scales well below about 100 nm [2]. We therefore impose local charge neutrality $z_+c_+ + z_-c_- = 0$ in our derivation exactly. This will allow us to identify the relevant measurable transport coefficients for the electrolytes used in Li ion batteries. For example, the strong Coulomb interaction between the ions prevent independent motion of ions to occur on the scale of battery cell dimensions. The main diffusion process will be correlated interdiffusion with a uniquely defined interdiffusion coefficient for positive and negative ions. Independent self diffusion of the different ions with

different self diffusion coefficient leading to slow charge separation is excluded in a strictly charge neutral system.

Under normal operation conditions for a Li ion battery we may safely assume that convection can be excluded as transport mechanism. This assumption allows to eliminate the concentration of the neutral solvent as independent variable. With M_0, M_+, M_- being the molar masses of solvent and positive and negative ions respectively we get in the absence of convection the relations

$$M_0 dc_0 + M_+ dc_+ + M_- dc_- = 0 \quad (3)$$

for the changes in the respective concentrations. Charge neutrality is then used to eliminate the concentration of the negative ions using the relation

$$z_+ c_+ + z_- c_- = 0 \quad (4)$$

It is therefore sufficient to determine the transport equations for the concentration $c = c_+ = -\frac{z_-}{z_+} c_-$. Using the constraints between changes in energy density u , entropy density s , concentration c and charge density q , and denoting as usually the temperature by T , the thermodynamic relation for the electrolyte in an external field Φ can be written as

$$du = Tds + \mu dc + \Phi dq \quad (5)$$

Due to the imposed charge neutrality the changes in the charge are zero i.e. $dq = 0$. The energy density also contains the contribution from the electric fields [14]. The effective chemical potential μ is a combination of the chemical potentials μ_+, μ_- and μ_0 of the ions and the solvent.

$$\mu = \tilde{\mu}_+ - \frac{z_+}{z_-} \tilde{\mu}_- \quad (6)$$

$$\tilde{\mu}_+ = \mu_+ - \frac{M_+}{M_0} \mu_0 \quad (7)$$

$$\tilde{\mu}_- = \mu_- - \frac{M_-}{M_0} \mu_0 \quad (8)$$

Formally, the chemical potential μ is the work to be performed for injecting 1mol of Li ions from infinity in the electrolyte including the work to rearrange the negative ions and neutral solvent molecules such that charge neutrality and momentum are conserved. The entropy production σ in the system fulfills the relation [12,13].

$$Td\sigma = -\mathbf{J}_s \nabla T - \mathbf{N}_+ \nabla \tilde{\mu}_+ - \mathbf{N}_- \nabla \tilde{\mu}_- - \mathbf{j} \nabla \Phi \quad (9)$$

The electric current is given by

$$\mathbf{j} = z_+ \mathbf{N}_+ + z_- \mathbf{N}_- \quad (10)$$

Using this relation to eliminate the flux of negative ions $\mathbf{N}_-$, we obtain

$$Td\sigma = -\mathbf{J}_s \nabla T - \mathbf{N}_+ \nabla \mu - \mathbf{j} \nabla \left(\Phi + \frac{\tilde{\mu}_-}{z_- F} \right) \quad (11)$$

where F is the Faraday number. Note that the form of the entropy production determines the set of independent thermodynamic forces and thus the correct form of the Onsager relations in the constitutive equations for the fluxes [12]. In the following we neglect for simplicity thermal fluctuations e.g. $dT = 0$.

Having identified the independent thermodynamic variables and forces, it is possible to formulate the constitutive relations for the fluxes. Under the necessary requirement of strictly positive entropy production they have the general form

$$\mathbf{N}_+ = -L_{11} \nabla \tilde{\mu}_+ - L_{12} \nabla \tilde{\Phi} \quad (12)$$

$$\mathbf{j} = -L_{21} \nabla \tilde{\mu}_+ - L_{22} \nabla \tilde{\Phi} \quad (13)$$

where $\tilde{\Phi} = \Phi + \frac{\tilde{\mu}_-}{z_- F}$. $\tilde{\Phi}$ may be interpreted as the renormalized effective potential due to the partial shielding of the external potential by the negative ions. The Onsager matrix L_{ij} has to be symmetric positive definite i.e. $L_{12} = L_{21}$. A simple rearrangement of (12), (13) and introduction of standard notation leads to

$$\mathbf{N}_+ = -D_e \nabla c + \frac{t_+}{F z_+} \mathbf{j} \quad (14)$$

$$\mathbf{j} = -\kappa \nabla \tilde{\Phi} - \kappa \frac{t_+}{F z_+} \left(\frac{\partial \tilde{\mu}_+}{\partial c} \right) \nabla c \quad (15)$$

The transport coefficients D_e, t_+, κ are the ion collective interdiffusion coefficient of the fully interacting system at zero electric current, the transference number and the ion conductivity, respectively. They are given by

$$\kappa = L_{22} \quad (16)$$

$$t_+ = \frac{z_+ L_{12} F}{\kappa} \quad (17)$$

$$D_e = (L_{11} - \kappa \frac{t_+^2}{F^2 z_+^2}) \left(\frac{\partial \tilde{\mu}_+}{\partial c} \right) = \frac{\det L}{L_{22}} \left(\frac{\partial \tilde{\mu}_+}{\partial c} \right) \quad (18)$$

The constitutive relation for the negative ion flux is a consequence of the definition of the current (10) and charge neutrality

$$\mathbf{N}_- = -D_e \nabla c_- + \frac{t_-}{F z_-} \mathbf{j} \quad (19)$$

Here $t_- = 1 - t_+$ is the transference number of negative ions. The interdiffusion coefficient for the density of negative ions is the same as the one for the positive ion density due to the imposed charge neutrality. This result is consistent with fundamental Green Kubo relation for the interdiffusion coefficient in a binary systems [15]. For comparison with experiments it is important to realize that it is the interdiffusion coefficient, and not the self diffusion coefficients, which has to be determined in order to simulate the behavior of Li ion batteries. In general the two self diffusion coefficients and the interdiffusion coefficient are mutually different from each other [15,16].

It is also important to note that the relation (15) is different from the one derived in [2]. The constitutive relation for the electrical current in [2] depends on the type of chemical reactions in the electrolyte and is not just a property of the local gradients in the independent field variables. This ansatz causes an asymmetry in the relations for the ion flux and the electrical current, which violates the fundamental Onsager relation necessary for strictly positive entropy production. In the case of a simple ion insertion reaction at the electrodes the factor t_+ in the relation for the current in (15), in [2] is replaced by $-(1-t_+)$. I.e. the absolute value and the sign in front of the ∇c term are different compared to our theory. The isothermal entropy production for the two models are

$$T\sigma_N = D_e \left(\frac{\partial \mu}{\partial c} \right)_T (\nabla c)^2 + \frac{j^2}{\kappa} - \left(\frac{\partial \mu}{\partial c} \right)_T \frac{j \nabla c}{F} \quad (20)$$

in the theory of [2] and

$$T\sigma_{LZ} = D_e \left(\frac{\partial \mu}{\partial c} \right)_T (\nabla c)^2 + \frac{j^2}{\kappa} \quad (21)$$

in our case. Since the thermodynamic derivative $\left(\frac{\partial \mu}{\partial c} \right)_T$ and the interdiffusion coefficient D_e are always positive the model presented here leads as expected to the strictly positive entropy production in Eq. (21). The last term in Eq. (20) does not have a definite sign and therefore allows in principal for negative entropy production.

Since the relation used in [2] is used as starting point for many battery modeling approaches [17,3,8], differences to our approach may be expected (cf Ref. [18]).

2.2 Transport in Active Particles

For the transport in the active particles, the diffusion and the conduction are essentially decoupled, since the mobility of the ions is much smaller than the one of the electrons and therefore the electric conduction is nearly completely carried by the electrons. The ions in the active particles are transported by diffusion only. The constitutive relations for ion flux and electrical current are given by

$$N_+ = -D_s \nabla c \quad (22)$$

$$j = -\sigma_s \nabla \Phi \quad (23)$$

where σ_s and D_s are the electronic conductivity and the ion diffusion coefficient respectively. As long as the binder and the additives in the electrodes are not treated as different phases the electronic conductivity is an effective conductivity of active particles and additives.

2.3 Intercalation Modeling and Interface Conditions

For the coupling of the transport in the active particles and in the solid electrolyte, interface conditions have to be formulated. The interface conditions describe the intercalation reaction and the de-intercalation reaction respectively on the mesoscopic scale (i.e. beyond the scale of the diffuse layer [2]). It is assumed that the transport of ions across the interface is completely described by the Butler Volmer expression i_{se} for the intercalation reaction [2].

$$i_{se} = i_0 \left(\exp\left[\frac{\alpha_a F}{RT} \eta_s\right] - \exp\left[\frac{-\alpha_c F}{RT} \eta_s\right] \right) \quad (24)$$

α_A and α_C with $\alpha_A + \alpha_C = 1$ are weighting the anodic and the cathodic contribution of the overpotential η_s to the overall reaction. A net current is flowing, if the electrochemical potential of electrolyte and active particle are not equal. The overpotential is the difference between the electrochemical potentials defined by

$$\eta_s := \Phi_s + \frac{\mu_s}{z_+ F} - \left(\Phi_e + \frac{\mu_e}{z_+ F} \right). \quad (25)$$

The chemical potential of the solid particle can be measured relative to the chemical potential of a Li metal electrode as half cell open circuit potential U_0

$$\mu_s = \mu_{Li} - z_+ F U_0 \quad (26)$$

Replacing μ_s in (25) by Eq. (26) gives

$$\eta_s := \Phi_s - \Phi_e - U_0 - \frac{\mu_e - \mu_{Li}}{z_+ F} \quad (27)$$

Usually the electrochemical potential φ_e is introduced with $\varphi_e = \Phi_e + \frac{\mu_e - \mu_{Li}}{z_+ F}$ and the overpotential is written as

$$\eta_s := \Phi_s - \varphi_e - U_0 \quad (28)$$

The amplitude i_0 in Eq. (24) is given by

$$i_0 = k c^{\alpha_a} c_s^{\alpha_a} \left(1 - \frac{c_s}{c_{s,max}} \right)^{\alpha_c} \quad (29)$$

k is a reaction rate. $c_{s,max}$ is the maximum concentration which can be stored in the active particle. We assume that Li ions are not stored in the double layer (i.e. all Li ions are intercalated in the active particle or released into the electrolyte). There should also be no flux of negative charges across the double layer. Neither enter electrons the electrolyte nor intercalate negative ions from the electrolyte in the active particles under ideal conditions. This especially means that the total current across the electrolyte-particle interface is due to the transport of positive ions only. If the particle is completely filled i.e. $c = c_{s,max}$, it has to be made sure by the interface conditions that no electrical current $\mathbf{j}$ is carried by

negative charge carriers across the interface. These conditions can be formulated mathematically in the following way with the normal $\mathbf{n}$ pointing from the solid into the electrolyte

$$\mathbf{j}_s \mathbf{n} = \mathbf{j}_e \mathbf{n} \quad (30)$$

$$\mathbf{j}_s \mathbf{n} = i_{se} \quad (31)$$

$$\mathbf{N}_{+,s} \mathbf{n} = \mathbf{N}_{+,e} \mathbf{n} \quad (32)$$

$$\mathbf{N}_{+,s} \mathbf{n} = \frac{i_{se}}{F} \quad (33)$$

To solve the model for the battery problem additional boundary conditions have to be provided for the potential and the current at the current collectors in contact with the active particles. These conditions are determined by the operating conditions of the battery. In addition the ion fluxes have to be set to zero at all external boundaries.

3 Conclusions

We derived a thermodynamically consistent model for transport of charges in a battery cell, consisting of active particles and electrolyte in cathode and anode. The final set of equations is given by (1) and (2), which have to be written down separately for anode, electrolyte and cathode. The respective fluxes for the electrolyte are given in (14), (15) and for the active particles in the cathode and the anode in (22) and (23). The transport coefficients for anode and cathode active particles are of course different. The interface conditions for the intercalation from the electrolyte in the active particle are formulated in (30) -(33) with i_{se} given in (24). We didn't formulate boundary conditions, since they depend on the details of the coupling of electrodes to some external electrical circuit. The modeling of the separator was not addressed, but it is straightforward using effective diffusion coefficients and ionic conductivities in the electrolyte theory, if the separator itself is a porous structure [1]. To test the model, a 1-D porous electrode version of the model was implemented in the commercial software package Comsol and compared with the model used in [1]. Detailed results will be presented in [18]. Numerical algorithm for the introduced model, as well as its numerical study for 3D geometry, are presented in [19]

Acknowledgment. The work was supported by the Fraunhofer system research for electromobility (FSEM) within the economic stimulus package II of the German Ministry of Education and Research.

References

1. Fuller, T.F., Doyle, M., Newman, J.: Simulation and optimization of the dual lithium ion insertion cell. J. Electrochem. Soc. 141, 1–10 (1994)
2. Newman, J., Thomas-Alyea, K.E.: Electrochemical Systems. Wiley, Chichester (2004)

3. Thomas, K.E., Newman, J., Darling, R.M.: Mathematical modeling of lithium batteries. In: Schalkwijk, W.A., Scrosati, B. (eds.) *Advances in Lithium-Ion Batteries*, pp. 345–392. Kluwer, Dordrecht (2002)
4. Botte, G.G., Subramanian, V.R., White, R.E.: Mathematical modeling of secondary lithium batteries. *Electrochimica Acta* 45, 2595–2609 (2000)
5. Danilov, D., Notten, P.H.L.: Mathematical modelling of ionic transport in the electrolyte of li-ion batteries. *Electrochimica Acta* 53, 5569–5578 (2008)
6. Olesen, L.H., Bazant, M.Z., Bruus, H.: Strongly nonlinear dynamics of electrolytes in large ac voltages. arXiv:0908.3501 (2009)
7. Wang, C.Y., Gu, W.B., Liaw, B.Y.: Micro-macroscopic coupled modeling of batteries and fuel cells. i. model development. *J. Electrochem. Soc.* 145, 3407–3417 (1998)
8. Wang, C.W., Sastry, A.M.: Mesoscale modeling of li-ion polymer cell. *J. Electrochem. Soc.* 154, A1035–A1047 (2007)
9. Zausch, J., Latz, A., Schmidt, S., Less, G.B., Seo, J.H., Han, S., Sastry, A.M.: Micro-scale modeling of li-ion batteries; parameterization and validation (2010) (to be published)
10. Vetter, J., Novak, P., Wagner, M.R., Veit, C., Möller, K.-C., Besenhard, J.O., Winter, M., Wohlfahrt-Mehrens, M., Vogler, C., Hammouche, A.: Ageing mechanisms in lithium-ion batteries. *J. Pow. Sources* 147, 269–281 (2005)
11. Landau, L.D., Lifshitz, E.M.: *Electrodynamics of Continuous Media*. Pergamon, Oxford (1984)
12. de Groot, S., Mazur, P.: *Non-Equilibrium Thermodynamics*. Dover, New York (1984)
13. Latz, A., Zausch, J.: Thermodynamic consistent transport theory of Li-Ion batteries. *J. Pow. Sources* (2010, in print)
14. Liu, M.: Hydrodynamic theory of electromagnetic fields in continuous media. *Phys. Rev. Lett.* 70, 3580–3583 (1993)
15. Hansen, J.P., McDonald, I.R.: *Theory of Simple Liquids*. Academic Press, London (1986)
16. Aouizerat-Elarby, A., Dez, H., Prevel, B., Jal, J., Bert, J., Dupuy-Philon, J.: Diffusion processes in LiCl, R H₂O solutions. *Journal of Molecular Liquids* 84(3), 289–299 (2000)
17. Doyle, M., Newman, J., Gozdz, A.S., Schmutz, C.N., Tarascon, J.M.: Comparison of modeling predictions with experimental data from plastic lithium ion cells. *J. Electrochem. Soc.* 143, 1890–1903 (1996)
18. Latz, A., Zausch, J.: Mesoscopic modeling and simulation of charge and ion transport in li ion battery cells. In: *Proceedings Dechema Conference on Materials for Energy* (2010)
19. Popov, P., Vutov, Y., Margenov, S., Iliev, O.: Finite volume discretization of nonlinear diffusion in li-ion batteries. In: Dimov, I., Dimova, S., Kolkovska, N. (eds.) *Numerical Methods and Applications*. LNCS, vol. 6064. Springer, Heidelberg (to appear)

Finite Volume Discretization of Equations Describing Nonlinear Diffusion in Li-Ion Batteries

P. Popov^{1,*}, Y. Vutov¹, S. Margenov¹, and O. Iliev²

¹ Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Sofia, Bulgaria
ppopov@parallel.bas.bg,

yavor.vutov@gmail.com, margenov@parallel.bas.bg

² Fraunhofer ITWM, D-67663 Kaiserslautern, Germany
iliev@itwm.fhg.de

Abstract. Numerical modeling of electrochemical process in Li-Ion battery is an emerging topic of great practical interest. In this work we present a Finite Volume discretization of electrochemical diffusive processes occurring during the operation of Li-Ion batteries. The system of equations is a nonlinear, time-dependent diffusive system, coupling the Li concentration and the electric potential. The system is formulated at length-scale at which two different types of domains are distinguished, one for the electrolyte and one for the active solid particles in the electrode. The domains can be of highly irregular shape, with electrolyte occupying the pore space of a porous electrode. The material parameters in each domain differ by several orders of magnitude and can be nonlinear functions of Li ions concentration and/or the electrical potential. Moreover, special interface conditions are imposed at the boundary separating the electrolyte from the active solid particles. The field variables are discontinuous across such an interface and the coupling is highly nonlinear, rendering direct iteration methods ineffective for such problems. We formulate a Newton iteration for a purely implicit Finite Volume discretization of the coupled system. A series of numerical examples are presented for different type of electrolyte/electrode configurations and material parameters. The convergence of the Newton method is characterized both as function of nonlinear material parameters and the nonlinearity in the interface conditions.

1 Introduction

The Li-Ion battery system is described mathematically as a coupled system of differential equations for the Li ions concentration, $c(\mathbf{x}, t)$, $\left[\frac{\text{mol}}{\text{cm}^3}\right]$ and the electric potential, $\phi(\mathbf{x}, t)$, $[V]$ in the domain Ω [3,2]. The domain is occupied

* Corresponding author.

by electrolyte and active particles. Their respective subdomains are denoted Ω_e and Ω_s , with $\Omega = \Omega_e \cup \Omega_s$ and $\Omega_e \cap \Omega_s = \emptyset$. The field equations are given by:

$$\frac{\partial c}{\partial t} - \nabla \cdot (\alpha(c, \phi) \nabla c + \beta(c, \phi) \nabla \phi) = 0 \quad \text{in } \Omega_s \text{ and } \Omega_e, \quad (1a)$$

$$-\nabla \cdot (\lambda(c, \phi) \nabla c + \kappa(c, \phi) \nabla \phi) = 0 \quad \text{in } \Omega_s \text{ and } \Omega_e, \quad (1b)$$

where $\kappa(c, \phi)$ is the ionic conductivity, a prescribed function, and:

$$\alpha(c, \phi) := \nu_+ D_e(c, \phi) + \frac{RT}{\nu_+ z_+ F^2} \frac{t_+(c) \kappa_D(c, \phi)}{c}, \quad \left[\frac{cm^2}{s} \right], \quad (2a)$$

$$\beta(c, \phi) := \kappa(c, \phi) \frac{t_+(c)}{\nu_+ z_+ F}, \quad \left[\frac{mol}{V \cdot cm \cdot s} \right], \quad (2b)$$

$$\lambda(c, \phi) := \frac{RT}{F} \frac{\kappa_D(c, \phi)}{c}, \quad \left[\frac{A \cdot cm^2}{mol} \right]. \quad (2c)$$

The dimensionless parameters $n = 1$, $s_+ = -1$, $z_+ = 1$, $z_- = -1$, $\nu_+ = \nu_- = 1$ indicate a single ionization state, F is the Faraday constant and R the gas constant. Next, κ_D is defined as follows:

$$\kappa_D(c, \phi) := \kappa(c, \phi) t_+(c, \phi). \quad (3)$$

a thermodynamic justification of this constitutive relationship is given in [1], together with an explanation of all the parameters. It should be noted that the model used is different from the classical work by Newman, [2,4], where:

$$\kappa_D(c, \phi) := \kappa(c, \phi) (\nu_+ + \nu_-) \left(\frac{s_+}{n\nu_+} + \frac{t_+(c)}{z_+\nu_+} - \frac{s_0 c}{nc_0} \right) \left(1 + \frac{\partial \ln f_+}{\partial \ln c} \right). \quad (4)$$

The model (3) has the advantage of being consistent with the entropy inequality for all possible thermodynamic loading paths. In contrast, (4) may generate negative entropy, a physically unacceptable situation [1].

The transference function t_+ allows us to distinguish between electrolyte and active particles. In an active particle, one has $t_+ \equiv 0$. In the electrolyte, t_+ is nonzero, typically an empirically measured function of c [4].

The system (1) is not complete without conditions on the interface $\Gamma = \partial\Omega_e \cap \partial\Omega_s$ between active particles and electrolyte. The flux of Li ions, which is implied by the model (1), is:

$$\mathbf{N} := -(\alpha(c, \phi) \nabla c + \beta(c, \phi) \nabla \phi), \quad (5)$$

and the flux of the electric potential, i.e. the current, is

$$\mathbf{J} := -\lambda(c, \phi) \nabla c + \kappa(c, \phi) \nabla \phi. \quad (6)$$

At the interface $\Gamma = \bar{\Omega}_e \cap \bar{\Omega}_s$ between a solid particle and electrolyte, one has a discontinuous concentration c and potential ϕ . We use subscript e and s to denote values on the interface when taken from the electrolyte side and from the

side of the active solid particles, respectively. The type of interface conditions to be imposed is subject to active research [2]. In this paper we follow [1], where two interface conditions, for each of the fluxes (5) and (6) are considered. One is that the normal component of each of the fluxes is continuous across an interface. Moreover, it is required that the value of the normal component of the flux is given by a nonlinear relationship of all the variables c_e , c_s , ϕ_e , ϕ_s , that is:

$$\mathbf{N}_s \mathbf{n} = \mathbf{N}_e \mathbf{n} = \mathcal{N}(c_e, c_s, \phi_e, \phi_s), \text{ on } \Gamma, \quad (7)$$

$$\mathbf{J}_s \mathbf{n} = \mathbf{J}_e \mathbf{n} = \mathcal{J}(c_e, c_s, \phi_e, \phi_s), \text{ on } \Gamma, \quad (8)$$

where the scalar functions $\mathcal{N}$ and $\mathcal{J}$ are defined as follows:

$$\eta = \phi_s - \phi_e - U_0 \quad (9)$$

$$\mathcal{J} = k \left(\frac{c_e}{c_e^0} \right)^{\alpha_a} \left(\frac{c_s}{c_s^0} \right)^{\alpha_a} \left(1 - \frac{c_s}{c_{s,max}} \right)^{\alpha_c} \left(\exp \left(\frac{\alpha_a F}{RT} \eta_s \right) - \exp \left(-\frac{\alpha_c F}{RT} \eta_s \right) \right) \quad (10)$$

$$\mathcal{N} = \frac{\mathcal{J}}{F}. \quad (11)$$

Note that when t_+ is constant in the electrolyte (it is always constant in the active particles), the divergence of the current is identically, zero, which allows to simplify the first equation in (1). As a result, the system (1) takes the following simplified form in either subdomain:

$$\frac{\partial c}{\partial t} - \nabla \cdot (\nu_+ D_e(c, \phi) \nabla c) = 0, \quad (12a)$$

$$-\nabla \cdot (\lambda(c, \phi) \nabla c + \kappa(c, \phi) \nabla \phi) = 0. \quad (12b)$$

If D_e is not a function of ϕ , the system (12) becomes completely decoupled in each subdomain. Note however, that the interface conditions (5)-(8) imply that the system is always coupled and always nonlinear, regardless of the coefficients.

2 Discretization

We present here the discretization for the general case, that is, the fully coupled system (1) is discretized by cell centered finite volumes. Let the domain Ω be partitioned into a polygonal mesh, e.g. $\Omega = \sum_{i=1}^N e_i$, with each cell e_i being a polygon/polyhedron. We suppose that the interface Γ does not cross any cell, instead, it is composed by cell faces. It is further required that this mesh is suitable for finite volume discretizations, that is, all vertices of e_i lie on a circle/sphere, whose center lies in the proper interior of e_i . By integrating the first equation over $e_i \times [t_n, t_{n+1}]$ and using the divergence theorem, one gets:

$$\begin{aligned} 0 &= \int_{t_n}^{t_{n+1}} \int_{e_i} \left(\frac{\partial c}{\partial t} - \nabla \cdot (\alpha(c, \phi) \nabla c + \beta(c, \phi) \nabla \phi) \right) dx dt \\ &= \int_{e_i} c(x, t_{n+1}) dx - \int_{e_i} c(x, t_n) dx - \int_{t_n}^{t_{n+1}} \int_{\partial e_i} (\alpha(c, \phi) \nabla c + \beta(c, \phi) \nabla \phi) \cdot \mathbf{n} dA. \end{aligned} \quad (13)$$

The second equation (1b) is similarly transformed as follows:

$$0 = - \int_{t_n}^{t_{n+1}} \int_{\partial e_i} (\lambda(c, \phi) \nabla c + \kappa(c, \phi) \nabla \phi) \cdot \mathbf{n} dA. \quad (14)$$

Now, denote by x_i the circumcenter of e_i and denote by $c_i(t)$ the value of the concentration at x_i , that is, $c_i(t) = c(x_i, t)$. Similarly, let $\phi_i(t) = \phi(x_i, t)$. The volume integral in (13) can be approximated by a one-point formula. Moreover, let e_j be a neighbor of e_i and denote by f_{ij} the face common to e_i and e_j . Denote by $\mathcal{N}_i$ the index set of all same type neighbors of e_i , that is, $\mathcal{N}_i = \{j \in \mathbb{N} | \bar{e}_j \cap \bar{e}_i = f_{ij} \neq \emptyset, \text{ and both } e_i \text{ and } e_j \text{ are either active particles or both are electrolyte}\}$. Using the standard midpoint flux approximation and assuming, for the time being, that e_i has no faces belonging to the interface Γ , one gets:

$$0 = |e_i| (c_i(t_{n+1}) - c_i(t_n)) - \int_{t_n}^{t_{n+1}} \sum_{j \in \mathcal{N}_i} |f_{ij}| \left(\alpha_{\frac{i+j}{2}} \frac{c_j(t) - c_i(t)}{d(x_i, x_j)} + \beta_{\frac{i+j}{2}} \frac{\phi_j(t) - \phi_i(t)}{d(x_i, x_j)} \right) dt, \quad (15)$$

$$0 = - \int_{t_n}^{t_{n+1}} \sum_{j \in \mathcal{N}_i} |f_{ij}| \left(\lambda_{\frac{i+j}{2}} \frac{c_j(t) - c_i(t)}{d(x_i, x_j)} + \kappa_{\frac{i+j}{2}} \frac{\phi_j(t) - \phi_i(t)}{d(x_i, x_j)} \right) dt, \quad (16)$$

where $\alpha_{\frac{i+j}{2}}$, $\beta_{\frac{i+j}{2}}$, $\lambda_{\frac{i+j}{2}}$, $\kappa_{\frac{i+j}{2}}$ are the harmonic averages of the respective coefficients at the midpoints of each face.

When a cell e_i has an interface face then (5) and (6) have to be incorporated into (15) and (16). Let e_i now share an interface face f_{ik} with e_k . Recall that $\mathcal{N}_i$ is defined as the index set of all *same type* neighbors, that is, $k \notin \mathcal{N}_i$. Thus, the flux corresponding to f_{ik} is missing from (15) and (16). Suppose, without loss of generality, that e_i is an electrolyte cell and e_k is occupied by solid. By approximating c and ϕ at Γ by the nearest cell values, the fluxes at f_{ij} become

$$\int_{t_n}^{t_{n+1}} |f_{ik}| \mathcal{N}(c_i(t), c_k(t), \phi_i(t), \phi_k(t)) dt, \quad (17)$$

$$\int_{t_n}^{t_{n+1}} |f_{ik}| \mathcal{J}(c_i(t), c_k(t), \phi_i(t), \phi_k(t)) dt. \quad (18)$$

These must be added to (15) and (16), respectively, for each interface face of e_i .

Next, we employ a backward Euler method to approximate the remaining time integrals. By denoting $C_i = c_i(t_{n+1})$ and $\Phi_i = \phi_i(t_{n+1})$ this results in the system of algebraic equations for $\mathbf{C}^{n+1}$, $\mathbf{\Phi}^{n+1}$:

$$0 = |e_i| \frac{C_i - c_i(t_n)}{dt} - \sum_{j \in \mathcal{N}_i} |f_{ij}| \left(\alpha_{\frac{i+j}{2}} \frac{C_j - C_i}{d(x_i, x_j)} + \beta_{\frac{i+j}{2}} \frac{\Phi_j - \Phi_i}{d(x_i, x_j)} \right) + \sum_{k \in \mathcal{I}_i} |f_{ik}| \mathcal{N}(C_i, C_k, \Phi_i, \Phi_k), \quad (19)$$

$$\begin{aligned}
0 = & - \sum_{j \in \mathcal{N}_i} |f_{ij}| \left(\lambda_{\frac{i+j}{2}} \frac{C_j - C_i}{d(x_i, x_j)} + \kappa_{\frac{i+j}{2}} \frac{\Phi_j - \Phi_i}{d(x_i, x_j)} \right) \\
& + \sum_{k \in \mathcal{I}_i} |f_{ik}| \mathcal{J}(C_i, C_k, \Phi_i, \Phi_k).
\end{aligned} \tag{20}$$

Here $\mathcal{I}_i$ is the set of cells that share an interface with e_i , and without loss of generality, e_i is an electrolyte cell. If e_i is a solid cell, then the sign of the interface fluxes has to be reversed.

3 Linearization

Due to the strong nonlinearities involved, the Newton method is used to linearize the system (19), (20) at each time step. Denote by $\mathbf{F}(\mathbf{C}, \Phi)$ and $\mathbf{G}(\mathbf{C}, \Phi)$ the right-hand sides of (19) and (20), respectively. The Newton iteration for the FV discretization of the (1) in component-wise form can be written as follows:

$$\begin{aligned}
0 = F_i(\mathbf{C}^{(k)}, \Phi^{(k)}) & + \sum_{j \in \mathcal{N}_i} \frac{\partial F_i}{\partial C_j}(\mathbf{C}^{(k)}, \Phi^{(k)}) (C_j^{(k+1)} - C_j^{(k)}) \\
& + \sum_{j \in \mathcal{N}_i} \frac{\partial F_i}{\partial \Phi_j}(\mathbf{C}^{(k)}, \Phi^{(k)}) (\Phi_j^{(k+1)} - \Phi_j^{(k)}),
\end{aligned} \tag{21}$$

$$\begin{aligned}
0 = G_i(\mathbf{C}^{(k)}, \Phi^{(k)}) & + \sum_{j \in \mathcal{N}_i} \frac{\partial G_i}{\partial C_j}(\mathbf{C}^{(k)}, \Phi^{(k)}) (C_j^{(k+1)} - C_j^{(k)}) \\
& + \sum_{j \in \mathcal{N}_i} \frac{\partial G_i}{\partial \Phi_j}(\mathbf{C}^{(k)}, \Phi^{(k)}) (\Phi_j^{(k+1)} - \Phi_j^{(k)}).
\end{aligned} \tag{22}$$

Computing the derivatives is straightforward. Assume, without loss of generality that e_l is the only interface neighbor to the electrolyte cell e_i . Then:

$$\begin{aligned}
\frac{\partial F_i}{\partial C_j} = & \frac{|e_i|}{dt} \delta_{ij} + \sum_{s \in \mathcal{N}_i} |f_{is}| \left[\alpha_{\frac{i+s}{2}}^{(k)} \frac{\delta_{sj} - \delta_{ij}}{d(x_i, x_s)} + \frac{\partial \alpha_{\frac{i+s}{2}}^{(k)}}{\partial C_j} \frac{C_s^{(k)} - C_i^{(k)}}{d(x_i, x_s)} + \frac{\partial \beta_{\frac{i+s}{2}}^{(k)}}{\partial C_j} \frac{\Phi_s^{(k)} - \Phi_i^{(k)}}{d(x_i, x_s)} \right] \\
& + |f_{il}| \left(\frac{\partial \mathcal{N}}{\partial C_e}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{ij} + \frac{\partial \mathcal{N}}{\partial C_s}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{lj} \right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial F_i}{\partial \Phi_j} = & \sum_{s \in \mathcal{N}_i} |f_{is}| \left[\beta_{\frac{i+s}{2}}^{(k)} \frac{\delta_{sj} - \delta_{ij}}{d(x_i, x_s)} + \frac{\partial \beta_{\frac{i+s}{2}}^{(k)}}{\partial \Phi_j} \frac{\Phi_s^{(k)} - \Phi_i^{(k)}}{d(x_i, x_s)} + \frac{\partial \alpha_{\frac{i+s}{2}}^{(k)}}{\partial \Phi_j} \frac{C_s^{(k)} - C_i^{(k)}}{d(x_i, x_s)} \right] \\
& + |f_{il}| \left(\frac{\partial \mathcal{N}}{\partial \Phi_e}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{ij} + \frac{\partial \mathcal{N}}{\partial \Phi_s}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{lj} \right),
\end{aligned}$$

where δ_{pq} is the Kroneker delta symbol. The expressions for the partial derivatives of $\mathbf{G}$ are similar:

$$\begin{aligned} \frac{\partial G_i}{\partial C_j} &= \sum_{s \in \mathcal{N}_i} |f_{is}| \left[\lambda_{\frac{i+s}{2}}^{(k)} \frac{\delta_{sj} - \delta_{ij}}{d(x_i, x_s)} + \frac{\partial \lambda_{\frac{i+s}{2}}^{(k)}}{\partial C_j} \frac{C_s^{(k)} - C_i^{(k)}}{d(x_i, x_s)} + \frac{\partial \kappa_{\frac{i+s}{2}}^{(k)}}{\partial C_j} \frac{\Phi_s^{(k)} - \Phi_i^{(k)}}{d(x_i, x_s)} \right] \\ &\quad + |f_{il}| \left(\frac{\partial \mathcal{J}}{\partial C_e}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{ij} + \frac{\partial \mathcal{J}}{\partial C_s}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{lj} \right), \\ \frac{\partial G_i}{\partial \Phi_j} &= \sum_{s \in \mathcal{N}_i} |f_{is}| \left[\kappa_{\frac{i+s}{2}}^{(k)} \frac{\delta_{sj} - \delta_{ij}}{d(x_i, x_s)} + \frac{\partial \kappa_{\frac{i+s}{2}}^{(k)}}{\partial \Phi_j} \frac{\Phi_s^{(k)} - \Phi_i^{(k)}}{d(x_i, x_s)} + \frac{\partial \lambda_{\frac{i+s}{2}}^{(k)}}{\partial \Phi_j} \frac{C_s^{(k)} - C_i^{(k)}}{d(x_i, x_s)} \right] \\ &\quad + |f_{il}| \left(\frac{\partial \mathcal{J}}{\partial \Phi_e}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{ij} + \frac{\partial \mathcal{J}}{\partial \Phi_s}(C_i^{(k)}, C_l^{(k)}, \Phi_i^{(k)}, \Phi_l^{(k)}) \delta_{lj} \right). \end{aligned}$$

The two field variables in our problems, c and ϕ , represent different physical quantities, which have very different scales. As a result, the stopping criteria for the Newton iteration has to be adjusted accordingly. A relative criterion was used individually for each component, that is, the iteration is terminated if:

$$\frac{\|\mathbf{F}(\mathbf{C}^{(k)}, \mathbf{\Phi}^{(k)})\|}{\|\mathbf{F}(\mathbf{C}^{(0)}, \mathbf{\Phi}^{(0)})\|} \leq TOL \text{ and } \frac{\|\mathbf{G}(\mathbf{C}^{(k)}, \mathbf{\Phi}^{(k)})\|}{\|\mathbf{G}(\mathbf{C}^{(1)}, \mathbf{\Phi}^{(1)})\|} \leq TOL \quad (23)$$

where TOL is a prescribed tolerance. Observe that the residual for the electrostatic equation (16) is scaled with the value at the first Newton iteration. The reason is the following. Given a converged time step t_n , the values for $\mathbf{c}(t_n)$ and $\phi(t_n)$ are used as initial guess for the Newton iteration for the time step t_{n+1} . However, the only difference in the residual will be contribution to $\mathbf{F}$ of the discretization of the time derivative in (15). Thus, the initial residual for $\mathbf{G}$ will be zero, rendering it useless for scaling purposes.

4 Numerical Examples

Two numerical examples are presented to test the model, the finite volume discretization, and the Newton algorithm. Both examples are on a micron length-scale, where the active particles and the electrolyte occupy distinctive domains.

Table 1. Material specific parameters and initial conditions

Material type	D_e	κ	c^0	c_{max}	U_0
	$\frac{cm^2}{s}$	$\left[\frac{A}{V \cdot cm}\right]$			
Electrolyte	7.5×10^{-7}	0.002	0.001		
Cathode	1.0×10^{-9}	0.038	0.020574	0.02286	0.001
Anode	3.9×10^{-10}	1.0	0.002639	0.02639	0

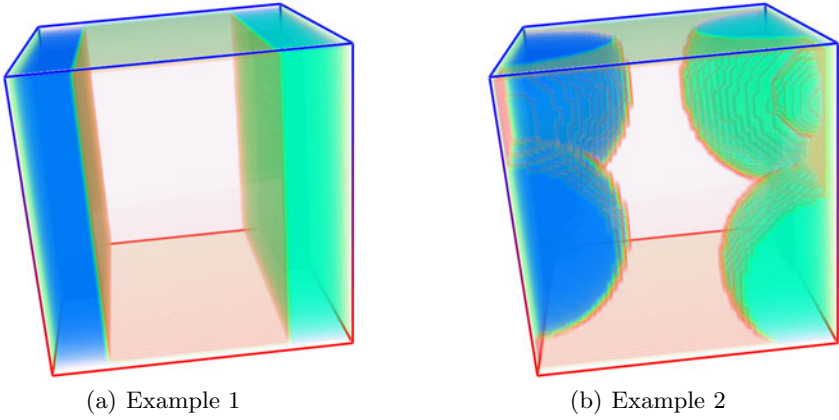


Fig. 1. Electrode geometry for each numerical example. The void space is occupied by the electrolyte.

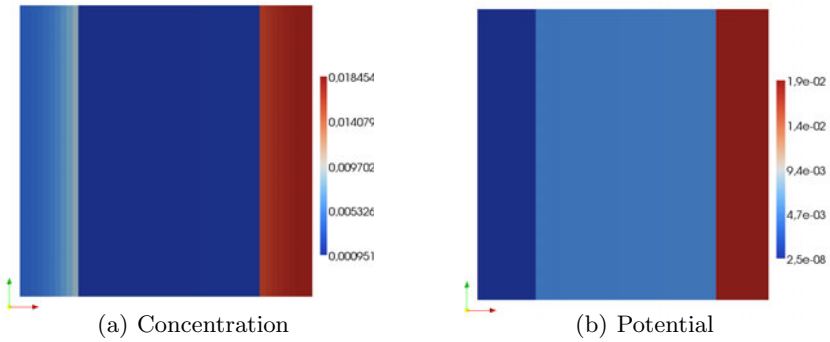


Fig. 2. Concentration (a) and potential (b) at time $t = 500s$ for the first example 1, $x - y$ cross-section

The geometry is given in Figure 1. In both cases, Ω is a cube with a $50\mu m$ side. The first example is a tests of a simple planar cathode-electrolyte-anode configuration. The second is representative of the porous microstructure of realistic active particles. Both examples are discretized on a 50^3 regular voxel grid.

The material constants and model parameters of (2) were taken as follows: $F = 96486 \left[\frac{A \cdot s}{mol} \right]$, $R = 8.314 \left[\frac{A \cdot V \cdot s}{K \cdot mol} \right]$ and $t_+(c) = 0.2$. The Li^+ diffusion coefficient D_e , ionic conductivity κ , the initial Li^+ concentrations c^0 , the maximum Li^+ concentration in the electrodes c_{max} and the open circuit potential for the electrodes U_0 , all material dependent parameters, are given in Table 1. All simulations were performed in isothermal conditions with $T = 300 [K]$.

The first series of numerical runs were performed with the above data. Since all material parameters were constant, the equations in each subdomain were linear, thus the nonlinearity was entirely due to the interface condition (7)-(11). The

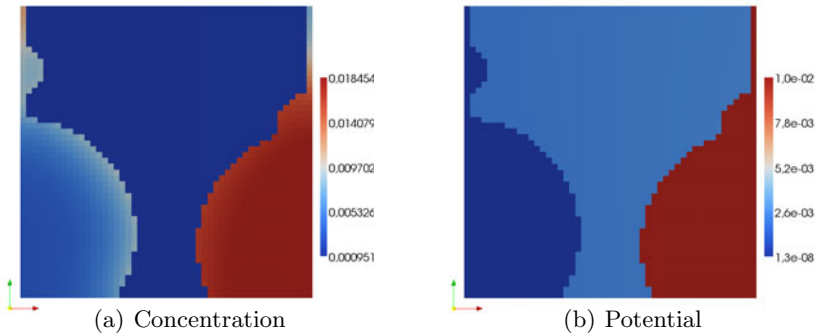


Fig. 3. Concentration (a) and potential (b) at time $t = 500s$ for the first example 1, $x - y$ cross-section

time step was 50s and a total of 20 steps were performed. It took slightly more than 1000s before the ionic concentration in parts of the domain became close to zero. A snapshot of the concentration and electric potential, for each of the two geometry examples, are given in Figures 2 and 3, respectively. Throughout the computational runs, the Newton iteration converged in 3 iterations at each time step, for both examples.

A second set of numerical experiments was performed, this time with nonlinear parameters for the electrolyte. In the absence of solid experimental data, a transference number $t_+ = 0.2 + 0.8c^2$ and $D_e = 1.27 \times 10^{-7}(1 + \phi^2)$ were used for the electrolyte, the remaining parameters being the same. These runs were done for the sake of testing the fully nonlinear system of equations. Again, the Newton iteration converged in 3 iterations at each time step, for both examples.

5 Conclusions

The main goal of this paper was to discretize and solve the system of coupled equations, which describes the diffusion of Li ions in a battery. A cell centered finite volume method was used to discretize the problem on a regular voxelized grid. The nonlinearity was treated with a full Newton method, both for the material parameters and the interface condition. It was found that the standard Newton method can handle both nonlinearities in a nearly optimal number of iterations.

Acknowledgments. The work was supported by the Fraunhofer system research for electromobility (FSEM) within the economic stimulus package II of the German Ministry of Education and Research. Peter Popov was also supported in part by EC grant FP7-PEOPLE-2007-4-3-IRG-230919 and US National Science Foundation grant NSF-DMS-0811180. Svetozar Margenov and Yavor Vutov were also supported in part by Bulgarian NSF GRANT DO 02-147/08.

References

1. Latz, A., Iliev, O., Zausch, J.: Modeling of species and charge transport in li-ion batteries. In: Proceedings of the Seventh International Conference on Numerical Methods and Applications, Borovets, Bulgaria, August 20-24 (2010)
2. Newman, J., Thomas-Alyea, K.E.: Electrochemical Systems. Wiley-Interscience, Hoboken (2004)
3. Thomas, K.E., Newman, J., Darling, R.M.: Mathematical modeling of lithium batteries, pp. 345–392. Kluwer Acad. Publ., Dordrecht (2002)
4. Wang, C., Sastry, A.M.: Mesoscale modeling of a li-ion polymer cell. Journal of The Electrochemical Society 154(11), A1035–A1047 (2007)

Numerical Study of Magnetic Flux in the LJJ Model with Double Sine-Gordon Equation

P.Kh. Atanasova, T.L. Boyadjiev, E.V. Zemlyanaya, and Yu.M. Shukrinov

JINR, Dubna, Russia

Abstract. The decrease of the barrier transparency in superconductor-insulator-superconductor (SIS) Josephson junctions leads to the deviations of the current-phase relation from the sinusoidal form. The sign of second harmonics is important for many applications, in particular in junctions with a more complex structure like SNINS or SFIFS, where N is a normal metal and F is a weak metallic ferromagnet. In our work we study the static magnetic flux distributions in long Josephson junctions taking into account the higher harmonics in the Fourier-decomposition of the Josephson current. Stability analysis is based on numerical solution of a spectral Sturm-Liouville problem formulated for each distribution. In this approach the nullification of the minimal eigenvalue of this problem indicates a bifurcation point in one of parameters. At each step of numerical continuation in parameters of the model, the corresponding nonlinear boundary problem is solved on the basis of the continuous analog of Newton's method. The solutions which do not exist in the traditional model have been found. The influence of second harmonic on stability of magnetic flux distributions for main solutions is investigated.

Keywords: long Josephson junction, in-line geometry, Sturm-Liouville, double sine-Gordon, bifurcation, continuous analog of Newton's method, fluxon, Numerov's finite-difference approximation.

1 Introduction

Physical properties of magnetic flux in Josephson junctions (JJs) deserve the base of the modern superconducting electronics. Tunnel SIS JJs are known to be having the sinusoidal current phase relation. However, the decrease of the barrier transparency in the SIS JJs leads the deviations of the current-phase relation from the sinusoidal form [1]. We study the static magnetic flux distributions in the long JJs taking into account the second harmonic in the Fourier-decomposition of the Josephson current. The sign of the second harmonic depends on physical applications under consideration. It is important, in particular, in junctions like SNINS and SFIFS, where N is a normal metal and F is a weak metallic ferromagnet [2]. Interesting properties of long Josephson junctions with an arbitrarily strong amplitude of second harmonic in current phase relation were considered in [3].

Our purpose was to investigate the effect of the second harmonic on the existence and stability of the magnetic flux distributions. Below, the numerical scheme and results of our stability analysis are demonstrated.

2 Mathematical Statement of the Problem

For a sufficiently wide class of JJ the superconducting Josephson current as a function of magnetic flux φ (phase difference of superconductors wave functions) can be represented as a sine series [4]:

$$I_S = I_c \sin \varphi + \sum_{m=2}^{\infty} I_m \sin m\varphi. \quad (1)$$

Using only first two terms of this expansion one can show [5] that the distribution of the magnitude $\varphi(x)$ along x -axis of the junction in the static regime [4] satisfies the double sine-Gordon equation (2SG).

$$-\varphi'' + a_1 \sin \varphi + a_2 \sin 2\varphi - \gamma = 0, \quad x \in (-l; l). \quad (2)$$

Here and below the prime means a derivative with respect to the coordinate x . The magnitude γ is the external current, l is the semilength of the junction, a_1 and a_2 are parameters corresponding to I_c and I_2 in (1) respectively. They depend on the preparation technology of junctions [1,6]. All the magnitudes are dimensionless.

In the case of in-line geometry of the junction the boundary conditions for (2) have the form

$$\varphi'(\pm l) = h_e, \quad (3)$$

where h_e is external magnetic field.

From the mathematical viewpoint the transfer of the junction into dynamical regime [4] means [7,8] a stability loss (bifurcation) of all static solutions $\varphi(x)$ of (2), (3) at the parameters γ or h_e variation. Our stability analysis of $\varphi(x, p)$ was based on numerical solution of the corresponding Sturm-Liouville problem

$$-\psi'' + q(x)\psi = \lambda\psi, \quad \psi'(\pm l) = 0 \quad (4)$$

with a potential $q(x) = a_1 \cos \varphi + 2a_2 \cos 2\varphi$.

The minimal eigenvalue $\lambda_0(p) > 0$ corresponds to a stable solution. In case $\lambda_0(p) < 0$ solution $\varphi(x, p)$ is unstable. The case $\lambda_0(p) = 0$ indicates the bifurcation with respect to one of parameters $p = (l, a_1, a_2, h_e, \gamma)$.

3 Numerical Method

Numerical solving of the boundary problem (2),(3) was performed on the basis of the Continuous analog of Newton's method [8]. At each Newtonian iteration the

corresponding linearized problem was solved using three-point Numerov's finite-difference approximation of the fourth order accuracy [9]. The discretization of the Sturm-Liouville problem (4) was realized with the help of standard second order finite-difference formulae. The calculation of the first several eigenvalues of the corresponding algebraic 3-diagonal problem was performed applying the standard subroutine from the package EISPACK. Details of numerical scheme are described in [10].

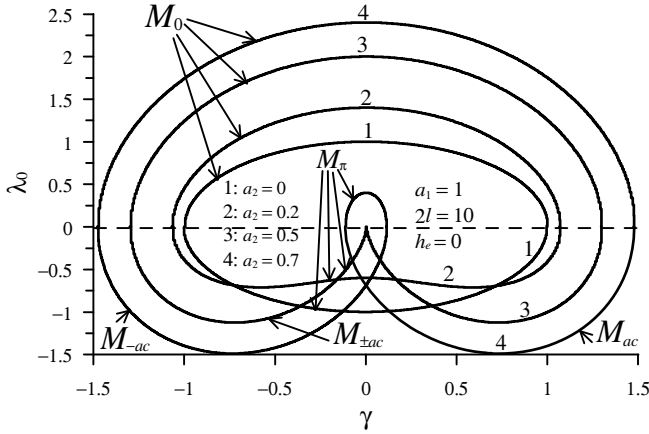


Fig. 1. Change of $\lambda_0(\gamma)$ for CS with increase of the coefficient a_2 in the interval $a_2 \in [0; 0.7]$ at $h_e = 0$, $a_1 = 1$, $2l = 10$

4 Numerical Results and Conclusions

Let us start with the *trivial solutions* of (2). In the “traditional” case $a_2 = 0$ two trivial solutions $\varphi = 0$ and $\varphi = \pi$ (below they are denoted by M_0 and M_π respectively) are known at $\gamma = 0$ and $h_e = 0$. Accounting of the second harmonic $a_2 \sin 2\varphi$ leads to appearing of two additional solutions $\varphi = \pm \arccos(-a_1/2a_2)$ (denoted as $M_{\pm ac}$). The corresponding λ_0 as functions of 2SG-equation coefficients have the form $\lambda_0[M_0] = a_1 + 2a_2$, $\lambda_0[M_\pi] = -a_1 + 2a_2$ and $\lambda_0[M_{\pm ac}] = (a_1^2 - 4a_2^2)/2a_2$. The exponential stability of these constant solutions (CS) is determined by the signs of the parameters a_1 , a_2 , and the ratio a_1/a_2 [10].

The dependencies of λ_0 on the external current γ for CS at several positive values of a_2 are demonstrated in Fig. 1. Arising of the stable states M_π by the external current γ at $a_2 > 0.5$ is shown.

When $a_2 < -0.5$ the stable solution M_0 disappears and other stable constant solutions $M_{\pm ac}$ arise. This transition is seen in Fig. 2.

In addition to CS, the 2SG equation admits *fluxon solutions*. The fluxons play a significant role in the JJ physics. Different distributions of magnetic flux in JJ are considered in the review [8]. At small external fields h_e such distributions are fluxon Φ^1 , antifluxon Φ^{-1} and their bound states $\Phi^1\Phi^{-1}$ and $\Phi^{-1}\Phi^1$. As external

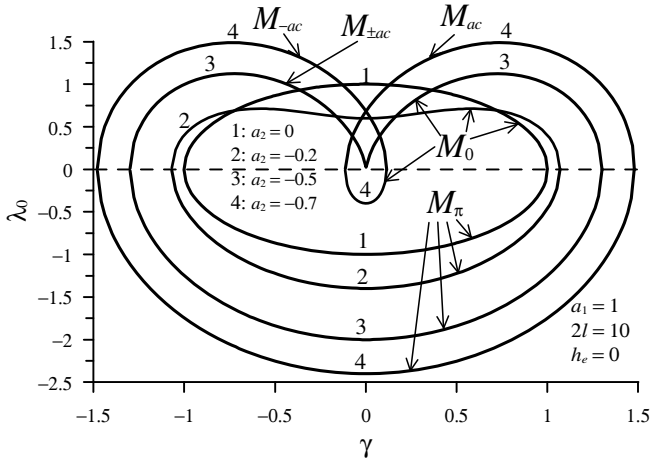


Fig. 2. The same as on Fig. 1 but for $a_2 \in [-0.7; 0]$

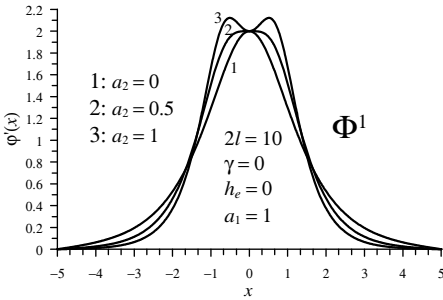


Fig. 3. Distribution of internal magnetic field of the fluxon Φ^1 for positive parameter a_2 at $\gamma = 0$, $h_e = 0$ and $2l = 10$

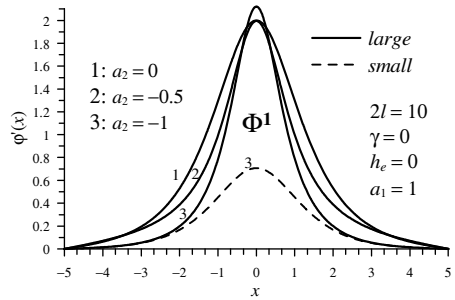


Fig. 4. The same as in Fig. 3 for negative a_2 . The dashed line shows “small” solution.

magnetic field h_e is growing, more complicated stable fluxon and bound states appear: $\Phi^{\pm n}$ and $\Phi^{\pm n} \Phi^{\mp n}$ ($n = 1, 2, 3, \dots$).

Let us compare some basic physical characteristics of one-fluxon solution Φ^1 in our model (2)–(3) with the traditional one ($a_1 = 1$, $a_2 = 0$). In both models the value of the magnetic flux $\varphi(x)$ in the middle of junction is $\varphi(0) = \pi$. In Fig. 3 the deformation of the $\varphi'(x)$ under influence of the parameter $a_2 \in [0; 1]$ is demonstrated. At $a_2 = 0.5$ the curve of internal magnetic field $\varphi'(x)$ has a plateau in a neighborhood of the center of junction $x = 0$. Further increase of the parameter a_2 leads to a formation of two maxima of the magnetic field. Thus, the inclusion of the second harmonic leads to the qualitative change of fluxon distribution Φ^1 . Such deformation does not appear with a decrease in parameter a_2 at $h_e = 0$ (Fig. 4). But, we observe a creation of new vortex when $a_2 < -0.5$ in zero magnetic field in agreement with the analytical results (see [3])

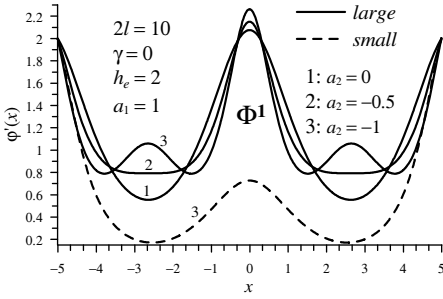


Fig. 5. The same as in Fig. 4 for $h_e = 2$

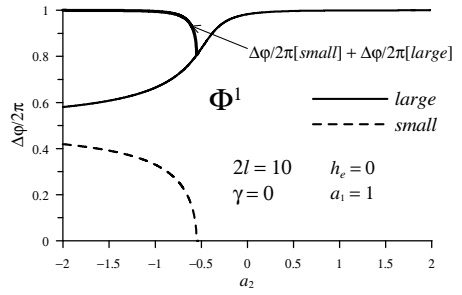


Fig. 6. Full magnetic flux for Φ^1 vs parameter $a_2 \in [-1; 1]$ at $h_e = 0$, $\gamma = 0$, $2l = 10$

and references there). This vortex is called as a “small” fluxon and coexisting fluxon solution as a “large” one. In cited work the new solution is investigated only at $h_e = 0$. In our work we show how the “small” fluxon is changed under the influence of the external magnetic field (Fig. 5). In the case of sufficiently large external magnetic field h_e a similar qualitative deformation is observed in the local minima regions only for the “large” fluxon when $a_2 < 0$ (see Fig. 5).

With change of the coefficient a_2 the number of fluxons [8]

$$N(p) = \frac{1}{2l\pi} \int_{-l}^l \varphi(x) dx,$$

corresponding to the “large” distribution Φ^1 is conserved, i.e., $\partial N / \partial a_2 = 0$. Here we have a value $N[\Phi^1] = 1$. But for the “small” vortex we have $N[\text{small}] = 0$, so in [11] we denote it as M_0 .

At $a_2 > -0.5$ the full magnetic flux [8] $\Delta\varphi(p) = \varphi(l) - \varphi(-l)$ for “large” fluxon solution tends to 2π when a_2 is growing. As we can see in Fig. 6, at $a_2 \lesssim -0.5$ $\Delta\varphi[\text{large}] + \Delta\varphi[\text{small}] \approx 2\pi$ except the region around the bifurcation value of the second harmonic $a_2 = -0.5$. So, due to this relation the creation of “large” and “small” fluxons might be considered as a one process. We consider that the creation of new solutions at $a_2 \lesssim -0.5$ and their relation with the traditional ones need a special investigation.

One-fluxon “large” state remains unstable in zero external magnetic field for all considered values of the parameter a_2 . The change of its stability under the influence of the field h_e is presented in [12].

In conclusion, we stress that new solutions we found do not exist in the traditional case ($a_2 = 0$). In this paper we focused on the stability analysis of constant and one-fluxon solutions only at different values of the a_2 . Investigation of another classes of solutions of 2GS-equation is a point of further research.

Acknowledgments. We thank to E. Goldobin for the stimulating discussions and important suggestions. The authors are thankful to I.V.Puzynin and T.P. Puzynina for valuable remarks and for the support of this work. The work of P.Kh. Atanasova is partially financed by the Program for collaboration of JINR-Dubna and Bulgarian scientific center “JINR – Bulgaria”. E.V. Zemlyanaya is grateful to RFFI (grant 09-01-00770-a) for a partial financial support.

References

1. Golubov, A.A., Kypriyanov, M.Yu., Il'ichev, E.: The current-phase relation in Josephson junctions. *Rev. Mod. Phys.* 76, 411–469 (2004)
2. Ryazanov, V.V., Oboznov, V.A., Rusanov, A.Yu., et al.: Coupling of two superconductors through a ferromagnet: evidence for a π junction. *Phys. Rev. Lett.* 36, 2427–2430 (2001)
3. Goldobin, E., Koelle, D., Kleiner, R., Buzdin, A.: Josephson junctions with second harmonic in the current-phase relation: Properties of junctions. *Phys. Rev. B* 76, 224523 (2007)
4. Likharev, K.K.: Introduction in Josephson junction dynamics. M. Nauka, GRFML (in Russian) (1985)
5. Hatakenaka, N., Takayanag, H., Kasai, Yo., Tanda, S.: Double sine-Gordon fluxons in isolated long Josephson junction. *Physica B* 284-288, 563–564 (2000)
6. Buzdin, A., Koshelev, A.E.: Periodic alternating 0-and π -junction structures as realization of φ -Josephson junctions. *Phys. Rev. B* 67, 220504(R) (2003)
7. Galpern, Yu.S., Filippov, A.T.: Joint solution states in inhomogeneous Josephson junctions. *Sov. Phys. JETP* 59, 894 (1984) (in Russian)
8. Puzynin, I.V., Boyadzhiev, T.L., Vinitskii, S.I., Zemlyanaya, E.V., Puzynina, T.P., Chuluunbaatar, O.: Methods of Computational Physics for Investigation of Models of Complex Physical Systems. *Physics of Particles and Nuclei*. 38(1), 70116 (2007)
9. Zemlyanaya, E.V., Puzynin, I.V., Puzynina, T.P.: PROGS2H4 – the software package for solving the boundary problem for the system of differential equations. *JINR Comm.* P11-97-414, Dubna, p. 18 (1997) (in Russian)
10. Atanasova, P.Kh., Zemlyanaya, E.V., Boyadjiev, T.L., Shukrinov, Yu.M.: Numerical modeling of long Josephson junctions in the frame of double sin-Gordon equation. *JINR Preprint* P11-2010-8, Dubna (2010); (accepted to *Journal of Mathematical modeling*)
11. Atanasova, P. Kh., Boyadjiev, T. L., Shukrinov, Yu. M., Zemlyanaya, E. V.: Influence of Josephson current second harmonic on stability of magnetic flux in long junctions, <http://arxiv.org/abs/1007.4778>
12. Atanasova, P. Kh., Boyadjiev, T. L., Shukrinov, Yu. M., Zemlyanaya, E. V.: Numerical investigation of the second harmonic effects in the LJ, <http://arxiv.org/abs/1005.5691>

A Simple Preconditioner for the SIPG Discretization of Linear Elasticity Equations

B. Ayuso¹, I. Georgiev², J. Kraus³, and L. Zikatanov⁴

¹ Centre de Recerca Matemàtica, Campus de Bellaterra
Edifici C, 08193 Bellaterra (Barcelona), Spain
bayuso@crm.cat

² Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences Acad. G. Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria
john@parallel.bas.bg

³ Johann Radon Institute for Computational and Applied Mathematics
Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria
johannes.kraus@oeaw.ac.at

⁴ Department of Mathematics, The Pennsylvania State University University Park,
PA 16802, USA
ltz@math.psu.edu

Abstract. We deal with the solution of the systems of linear algebraic equations arising from Symmetric Interior Penalty discontinuous Galerkin (SIPG) discretization of linear elasticity problems in primal (displacement) formulation. The main focus of the paper is on constructing a uniform preconditioner which is based on a natural splitting of the space of piecewise linear discontinuous functions. The presented approach has recently been introduced in [2] in the context of designing subspace correction methods for scalar elliptic partial differential equations and is extended here to linear elasticity equations, i.e., a class of vector field problems. Similar to the scalar case the solution of the linear algebraic system corresponding to the SIPG method is reduced to the solution of a problem arising from discretization by nonconforming Crouzeix-Raviart elements plus the solution of a well-conditioned problem on the complementary space.

1 Introduction

Let $\Omega \subset \mathbb{R}^2$ be a convex polygon and let $\mathbf{u}$ be a vector field in $\mathbb{R}^2$, defined on a domain Ω such that $\mathbf{u} \in [H^2(\Omega)]^2$. We also denote by $\langle \cdot, \cdot \rangle$ the Euclidean (resp. by $\langle \cdot : \cdot \rangle$ the Frobenius) scalar product for two vectors (resp. tensors) in $\mathbb{R}^2$ (resp. $\mathbb{R}^{2 \times 2}$), i.e.,

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{k=1}^2 v_k w_k, \quad \langle \mathbf{v} : \mathbf{w} \rangle = \sum_{j=1}^2 \sum_{k=1}^2 v_{jk} w_{jk}.$$

The corresponding products in $[L_2(\Omega)]^2$ and $[L_2(\Omega)]^{2 \times 2}$ are

$$(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \langle \mathbf{v}, \mathbf{w} \rangle, \quad (\mathbf{v} : \mathbf{w}) = \int_{\Omega} \langle \mathbf{v} : \mathbf{w} \rangle.$$

We denote by $\varepsilon(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$ the symmetric part of the gradient of $\mathbf{u}$. Consider now the linear elasticity problem: Find the displacement field $\mathbf{u}$ and the symmetric stress tensor $\boldsymbol{\sigma}$ such that:

$$\begin{aligned}\boldsymbol{\sigma} &= \lambda \operatorname{div} \mathbf{u} \mathbf{I} + 2\mu \varepsilon(\mathbf{u}) \quad \text{on } \Omega \\ -\operatorname{div} \boldsymbol{\sigma} &= \mathbf{f}, \quad \text{on } \Omega, \\ \boldsymbol{\sigma} \mathbf{n} &= \mathbf{0}, \quad \text{on } \Gamma_N, \\ \mathbf{u} &= \mathbf{g}, \quad \text{on } \Gamma_D.\end{aligned}\tag{1}$$

Here $\mathbf{u}$ takes prescribed values on a closed part of the boundary Γ_D (Dirichlet boundary) and satisfies natural (traction free) boundary conditions on the rest of the boundary Γ_N and $\mathbf{n}$ is the outward unit normal vector on $\partial\Omega$. Furthermore, λ and μ are the Lamé constants, satisfying $0 < \mu_1 < \mu < \mu_2$ and $0 < \lambda < \infty$, where incompressible behavior is obtained when $\lambda \rightarrow \infty$.

Let $\mathcal{T}_h$ be a shape-regular triangulation of Ω . We denote by h_T the diameter of the triangle T and we set $h = \max_{T \in \mathcal{T}_h} h_T$. A face (shared by two neighboring elements or being part of the boundary) is denoted by E . We denote the set of all faces by $\mathcal{E}_h$, and the collection of all interior faces and boundary faces by $\mathcal{E}_h^o$ and $\mathcal{E}_h^\partial$, respectively. Further, the set of Dirichlet faces is denoted by $\mathcal{E}_h^D$, and the set of Neumann faces by $\mathcal{E}_h^N$. We thus have,

$$\mathcal{E}_h = \mathcal{E}_h^o \cup \mathcal{E}_h^\partial, \quad \mathcal{E}_h^D = \mathcal{E}_h^\partial \cap \Gamma_D, \quad \mathcal{E}_h^N = \mathcal{E}_h^\partial \cap \Gamma_N, \quad \mathcal{E}_h^\partial = \mathcal{E}_h^D \cup \mathcal{E}_h^N.$$

For two vector fields $\mathbf{v}$ and $\mathbf{w}$, which are sufficiently smooth so that the integrals below exist, we denote

$$(\mathbf{v}, \mathbf{w})_{\mathcal{T}_h} = \sum_{T \in \mathcal{T}_h} \int_T \langle \mathbf{v}, \mathbf{w} \rangle, \quad (\mathbf{v}, \mathbf{w})_{\mathcal{E}} = \sum_{E \in \mathcal{E}} \int_E \langle \mathbf{v}, \mathbf{w} \rangle,$$

where $\mathcal{E} \subset \mathcal{E}_h$.

To define the average and jump trace operators for an interior face $E \in \mathcal{E}_h^o$, and any $T \in \mathcal{T}_h$, such that $E \in \partial T$ we set $\mathbf{n}_{E,T}$ to be the unit outward (with respect to T) normal vector to E . With every face $E \in \mathcal{E}_h^o$ we also associate a unit vector $\mathbf{n}_E$ which is orthogonal to E . For the boundary faces, we always set $\mathbf{n}_E = \mathbf{n}_{E,T}$, where T is the *unique* element for which we have $E \subset \partial T$. In our setting, for the interior faces, the particular direction of $\mathbf{n}_E$ is not important, although it is important that this direction is fixed. For every face $E \in \mathcal{E}_h$, we define $T^+(E)$ and $T^-(E)$ as follows:

$$\begin{aligned}T^+(E) &:= \{T \in \mathcal{T}_h \text{ such that } E \subset \partial T, \text{ and } \langle \mathbf{n}_E, \mathbf{n}_{E,T} \rangle > 0\}, \\ T^-(E) &:= \{T \in \mathcal{T}_h \text{ such that } E \subset \partial T, \text{ and } \langle \mathbf{n}_E, \mathbf{n}_{E,T} \rangle < 0\}.\end{aligned}\tag{2}$$

It is clear that for every face we have exactly one $T^+(E)$ and for the interior faces we also have exactly one $T^-(E)$. In the following, we will also write $T^\pm$ instead of $T^\pm(E)$. For a given function $\mathbf{w} \in [L_2(\Omega)]^2$ and a fixed interior face $E \in \mathcal{E}_h^o$

the average and jump trace operators are defined by $\{\!\!\{ \mathbf{w} \}\!\!\} := (\mathbf{w}^+ + \mathbf{w}^-)/2$, $\llbracket \mathbf{w} \rrbracket := (\mathbf{w}^+ - \mathbf{w}^-)$, where $\mathbf{w}^+$ and $\mathbf{w}^-$ denote the traces of $\mathbf{w}$ onto E taken from within the interior of T^+ and T^- , respectively. On boundary faces $E \in \mathcal{E}_h^\partial$, we set $\{\!\!\{ \mathbf{w} \}\!\!\} = \mathbf{w}$ and $\llbracket \mathbf{w} \rrbracket = \mathbf{w}$.

The space of piecewise smooth functions and the linear DG space (space of piecewise linear discontinuous functions) are defined by

$$[H^2(\mathcal{T}_h)]^2 = \{ \mathbf{u} \in [L^2(\Omega)]^2 \text{ such that } \mathbf{u}|_T \in [H^2(T)]^2, \quad \forall T \in \mathcal{T}_h \},$$

$$\mathbf{V}^{\text{DG}} := \{ u \in L^2(\Omega) \text{ such that } u|_T \in \mathbb{P}^1(T), \quad \forall T \in \mathcal{T}_h \},$$

where $\mathbb{P}^1(T)$ is the space of linear functions on T . The corresponding space of vector valued functions is then $\mathbf{V}^{\text{DG}} := [\mathbf{V}^{\text{DG}}]^2$.

The weak formulation of the linear elasticity problem reads as follows: Find $\mathbf{u} \in [H^2(\mathcal{T}_h)]^2$ such that

$$\mathcal{A}(\mathbf{u}, \mathbf{w}) = \mathcal{F}(\mathbf{w}), \quad \forall \mathbf{w} \in [H^2(\mathcal{T}_h)]^2. \quad (3)$$

Following [8], the bilinear form $\mathcal{A}(\cdot, \cdot)$ is given by

$$\mathcal{A}(\mathbf{u}, \mathbf{w}) = \mathcal{A}_0(\mathbf{u}, \mathbf{w}) + a_{j,1}(\llbracket \mathbf{u} \rrbracket, \llbracket \mathbf{w} \rrbracket), \quad (4)$$

where

$$\begin{aligned} \mathcal{A}_0(\mathbf{u}, \mathbf{w}) = & (\mathcal{C}\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{w}))_{\mathcal{T}_h} - (\{\!\!\{ (\mathcal{C}\varepsilon(\mathbf{u}))\mathbf{n} \}\!\!\}, \llbracket \mathbf{w} \rrbracket)_{\mathcal{E}_h} \\ & - (\llbracket \mathbf{u} \rrbracket, \{\!\!\{ (\mathcal{C}\varepsilon(\mathbf{w}))\mathbf{n} \}\!\!\})_{\mathcal{E}_h} + a_{j,0}(\llbracket \mathbf{u} \rrbracket, \llbracket \mathbf{w} \rrbracket). \end{aligned} \quad (5)$$

Here we set

$$\begin{aligned} a_{j,0}(\llbracket \mathbf{u} \rrbracket, \llbracket \mathbf{v} \rrbracket) &:= \alpha_0 \beta_0 \sum_{E \in \mathcal{E}_h} \int_E \langle h_E^{-1} \llbracket \mathbf{u} \rrbracket, \mathcal{P}_E^0 \llbracket \mathbf{v} \rrbracket \rangle, \\ a_{j,1}(\llbracket \mathbf{u} \rrbracket, \llbracket \mathbf{v} \rrbracket) &:= \alpha_1 \beta_1 \sum_{E \in \mathcal{E}_h} \int_E \langle h_E^{-1} \llbracket \mathbf{u} \rrbracket, \llbracket \mathbf{v} \rrbracket \rangle. \end{aligned} \quad (6)$$

The parameters α_i, β_i , $i = 0, 1$, are chosen so that the resulting SIPG discretization is consistent and stable, cf. [8]. The parameters β_0 and β_1 in (6) depend on the Lamé constants λ and μ and are $\beta_0 := 2(\lambda + \mu)$, $\beta_1 := 2\mu$. The remaining two parameters, α_0 and α_1 are at our disposal and they can serve to obtain different schemes.

Finally, to obtain the discrete formulation, we replace $[H^2(\mathcal{T}_h)]^2$ in (3) by $\mathbf{V}^{\text{DG}}$, and hence get the discrete problem: Find $\mathbf{u}_h \in \mathbf{V}^{\text{DG}}$ such that

$$\mathcal{A}(\mathbf{u}_h, \mathbf{w}) = \mathcal{F}(\mathbf{w}), \quad \forall \mathbf{w} \in \mathbf{V}^{\text{DG}}. \quad (7)$$

As we mentioned earlier the discretization that we introduced is exactly the SIPG discretization for the elasticity system introduced in [8].

2 Preconditioning

Let us introduce the classical Crouzeix-Raviart finite element space

$$V^{\text{CR}} = \{v \in L^2(\Omega) : v|_T \in \mathbb{P}^1(T), \forall T \in \mathcal{T}_h \text{ and } \mathcal{P}_E^0[v] = 0, \forall E \in \mathcal{E}_h^o\}. \quad (8)$$

Here for a given face E , the operator $\mathcal{P}_E^0 : L^2(E) \mapsto \mathbb{P}^0(E)$ denotes the L^2 -projection onto the constant function on E defined (for both scalar and vector valued functions) by $\mathcal{P}_E^0 \mathbf{w} = \frac{1}{|E|} \int_E \mathbf{w}$, for all $\mathbf{w} \in [L^2(E)]^2$. The corresponding space of vector valued functions is

$$\mathbf{V}^{\text{CR}} := [V^{\text{CR}}]^2. \quad (9)$$

Following [2] we introduce also the space complementary to V^{CR} in V^{DG} ,

$$\mathcal{Z} = \{z \in L^2(\Omega) : z|_T \in \mathbb{P}^1(T) \forall T \in \mathcal{T}_h \text{ and } \mathcal{P}_E^0[z] = 0 \forall E \in \mathcal{E}_h^o\}. \quad (10)$$

The corresponding space of vector valued functions is

$$\mathcal{Z} = \mathcal{Z}^2. \quad (11)$$

To describe the basis functions associated with the spaces (9) and (11), let $\varphi_{E,T}$ denote the canonical scalar Crouzeix-Raviart (CR) basis function on T , dual to the degree of freedom at the mass center of the face E , and extended as zero outside T . For $E \in \partial T$, $E' \in \partial T$, the function $\varphi_{E,T}$ satisfies

$$\varphi_{E,T}(m_{E'}) = \begin{cases} 1 & \text{if } E = E', \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we have $\varphi_{E,T} \in \mathbb{P}^1(T)$, and $\varphi_{E,T}(x) = 0$ for all $x \notin T$.

We observe that any function $\mathbf{u} \in \mathbf{V}^{\text{DG}}$ can be represented as

$$\begin{aligned} \mathbf{u}(x) &= \sum_{T \in \mathcal{T}_h} \sum_{E \in \partial T} \mathbf{u}_T(m_E) \varphi_{E,T}(x) \\ &= \sum_{E \in \mathcal{E}_h} \mathbf{u}^+(m_E) \varphi_E^+(x) + \sum_{E \in \mathcal{E}_h^o} \mathbf{u}^-(m_E) \varphi_E^-(x), \end{aligned} \quad (12)$$

where in the last identity we changed the order of summation and used the short hand notation $\varphi_E^\pm(x) := \varphi_{E,T^\pm}(x)$ together with

$$\begin{aligned} \mathbf{u}^\pm(m_E) &:= \mathbf{u}_{T^\pm}(m_E) = \frac{1}{|E|} \int_E \mathbf{u}^\pm ds, & \forall E \in \mathcal{E}_h^o, : E = \partial T^+ \cap \partial T^-, \\ \mathbf{u}(m_E) &:= \mathbf{u}_T(m_E) = \frac{1}{|E|} \int_E \mathbf{u}_T ds, & \forall E \in \mathcal{E}_h^\partial, \text{ such that } E = \partial T \cap \partial \Omega. \end{aligned}$$

We recall the definitions of $T^+(E)$ and $T^-(E)$ (see equation (2)) and set

$$\begin{aligned} \varphi_E^{\text{CR}} &= \varphi_{E,T^+(E)} + \varphi_{E,T^-(E)}, & \forall E \in \mathcal{E}_h^o, \\ \varphi_E^{\text{CR}} &= \varphi_{E,T^+(E)}, & \forall E \in \mathcal{E}_h^N. \end{aligned} \quad (13)$$

and

$$\begin{aligned}\psi_E^z &= \frac{\varphi_{E,T^+(E)} - \varphi_{E,T^-(E)}}{2}, & \forall E \in \mathcal{E}_h^o, \\ \psi_E^z &= \varphi_{E,T^+(E)}, & \forall E \in \mathcal{E}_h^D.\end{aligned}\quad (14)$$

Clearly, $\{\varphi_E^{CR}\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^N}$ are linearly independent, and $\{\psi_{E,T}^z\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^D}$ are linearly independent. A simple calculation then shows that

$$\mathbf{V}^{CR} = \text{span} \left\{ \{\varphi_E^{CR} \mathbf{e}_k\}_{k=1}^d \right\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^N}, \quad \mathbf{Z} = \text{span} \left\{ \{\psi_E^z \mathbf{e}_k\}_{k=1}^d \right\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^D}.$$

Here $\mathbf{e}_k$, $k = 1, \dots, d$ is the k -th canonical basis vector in $\mathbb{R}^d$. Hence by performing a change of basis in (12), we have obtained a “natural” splitting

$$\mathbf{V}^{DG} = \mathbf{V}^{CR} \oplus \mathbf{Z}$$

where the set $\{\psi_{E,T}^z\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^D} \cup \{\varphi_E^{CR}\}_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^N}$ provides a natural basis for the linear DG space. This is summarized in the next proposition (cf. [1]).

Proposition 1. *For any $\mathbf{u} \in \mathbf{V}^{DG}$ there exist unique $\mathbf{v} \in \mathbf{V}^{CR}$ and a unique $\mathbf{z} \in \mathbf{Z}$ such that*

$$\mathbf{u} = \mathbf{v} + \mathbf{z} \quad \text{and} \quad \begin{aligned} \mathbf{v} &= \sum_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^N} \left(\frac{1}{|E|} \int_E \llbracket \mathbf{u} \rrbracket ds \right) \varphi_E^{CR}(x) \in \mathbf{V}^{CR}, \\ \mathbf{z} &= \sum_{E \in \mathcal{E}_h^o \cup \mathcal{E}_h^D} \left(\frac{1}{|E|} \int_E \llbracket \mathbf{u} \rrbracket ds \right) \psi_{E,T^+}^z(x) \in \mathbf{Z}. \end{aligned}\quad (15)$$

The following property of the decomposition (15) suggests the construction of a subspace correction method.

Lemma 1. *Let $\mathbf{u} \in \mathbf{V}^{DG}$ be such that $\mathbf{u} = \mathbf{v} + \mathbf{z}$ with $\mathbf{v} \in \mathbf{V}^{CR}$ and $\mathbf{z} \in \mathbf{Z}$. Let $\mathcal{A}_0(\cdot, \cdot)$ be the bilinear form defined in (5). Then,*

$$\mathcal{A}_0(\mathbf{v}, \mathbf{z}) = \mathcal{A}_0(\mathbf{z}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathbf{V}^{CR}, \quad \forall \mathbf{z} \in \mathbf{Z}. \quad (16)$$

Hence the decomposition (15) is $\mathcal{A}_0$ -orthogonal, i.e., $\mathbf{V}^{CR} \perp_{\mathcal{A}_0} \mathbf{Z}$.

Using Equations (4)–(6), we find that for any $\mathbf{u}, \mathbf{w} \in \mathbf{V}^{DG}$, we can write $\mathbf{u} = \mathbf{z} + \mathbf{v}$, and $\mathbf{w} = \psi + \varphi$, where $\mathbf{z}, \psi \in \mathbf{Z}$ and $\mathbf{v}, \varphi \in \mathbf{V}^{CR}$, such that the bilinear form becomes $\mathcal{A}(\mathbf{u}, \mathbf{w}) = \mathcal{A}((\mathbf{z}, \mathbf{v}), (\psi, \phi))$. A simple calculation shows that

$$\mathcal{A}_0((\mathbf{z}, \mathbf{v}), (\psi, \phi)) = \mathcal{A}_0(\mathbf{z}, \psi) + \mathcal{A}_0(\mathbf{v}, \phi).$$

While in the scalar case it is possible to use $\mathcal{A}_0$ as a preconditioner of $\mathcal{A}$ (see [2]), in general, this is not a proper choice for the elasticity problem. In the latter case, however, a reasonable approximation of $\mathcal{A}(\cdot, \cdot)$ is given by the following block-diagonal preconditioner

$$\mathcal{B}((\mathbf{z}, \mathbf{v}), (\psi, \phi)) := \mathcal{A}(\mathbf{z}, \psi) + \mathcal{A}(\mathbf{v}, \phi). \quad (17)$$

Remark 1. Note that for traction free boundary conditions, $\mathcal{A}_0(\cdot, \cdot)$ is not equivalent to $\mathcal{A}(\cdot, \cdot)$ (see [6]), and in fact, even for bounded values of the Lamé constant λ the restriction of $\mathcal{A}_0(\cdot, \cdot)$ on $\mathbf{V}^{CR}$ is singular and does not satisfy the discrete analogue of the Korn's inequality.

The following algorithm describes the application of a preconditioner, which is based on the bilinear form in Equation (17).

Algorithm 1. *Let $\mathbf{r} \in [L_2(\Omega)]^2$ be given. Then the action of the preconditioner on $\mathbf{r}$ is the function $\mathbf{u} \in \mathbf{V}^{DG}$ which is obtained from the following three steps.*

1. Find $\mathbf{z} \in \mathcal{Z}$ such that $\mathcal{A}(\mathbf{z}, \boldsymbol{\psi}^z) = (\mathbf{r}, \boldsymbol{\psi}^z)_{T_h}$ for all $\boldsymbol{\psi}^z \in \mathcal{Z}$.
2. Find $\mathbf{v} \in \mathbf{V}^{CR}$ such that $\mathcal{A}(\mathbf{v}, \boldsymbol{\varphi}) = (\mathbf{r}, \boldsymbol{\varphi})_{T_h}$ for all $\boldsymbol{\varphi} \in \mathbf{V}^{CR}$.
3. Set $\mathbf{u} = \mathbf{z} + \mathbf{v}$.

The main result, which is formulated in Theorem 2 below, is that this algorithm provides a uniform preconditioner for $\mathcal{A}(\cdot, \cdot)$. The following lemma is crucial for obtaining this result. For the proofs of Lemma 2 and Theorem 2 we refer the reader to [1].

Lemma 2. *The following inequality holds for any $\mathbf{z} \in \mathcal{Z}$ and any $\mathbf{v} \in \mathbf{V}^{CR}$*

$$\mathcal{A}(\mathbf{z}, \mathbf{v})^2 \leq \gamma^2 \mathcal{A}(\mathbf{z}, \mathbf{z}) \mathcal{A}(\mathbf{v}, \mathbf{v})$$

where the constant $\gamma < 1$ is uniformly bounded.

Remark 2. Note that γ is uniformly bounded, which means that $\gamma \leq q < 1$ holds independently of the mesh size h and of the Lamé parameters λ and μ for some constant $q < 1$.

The next Theorem shows that the preconditioner given by Algorithm 1 is uniform with respect to the mesh size and the problem parameters.

Theorem 2. *Let $\mathcal{A}(\cdot, \cdot)$ be the bilinear form defined by (4) and $\mathcal{B}(\cdot, \cdot)$ be the bilinear form defined by (17). Then the following estimates hold for all $\mathbf{z} \in \mathcal{Z}$ and for all $\mathbf{v} \in \mathbf{V}^{CR}$*

$$\frac{1}{1+\gamma} \mathcal{A}((\mathbf{z}, \mathbf{v}), (\mathbf{z}, \mathbf{v})) \leq \mathcal{B}((\mathbf{z}, \mathbf{v}), (\mathbf{z}, \mathbf{v})) \leq \frac{1}{1-\gamma} \mathcal{A}((\mathbf{z}, \mathbf{v}), (\mathbf{z}, \mathbf{v})). \quad (18)$$

Here γ is the same constant that appears in Lemma 2.

3 Numerical Examples

We consider the model problem (1) with mixed boundary conditions on an L-shaped domain Ω (see Figure 1). We have taken for the penalty parameters in (6) the values $\alpha_0 = 4$ and $\alpha_1 = 1$. The initial triangulation (level 0) consists of 38 triangles. Each refinement level ℓ is obtained by subdividing each of the triangles

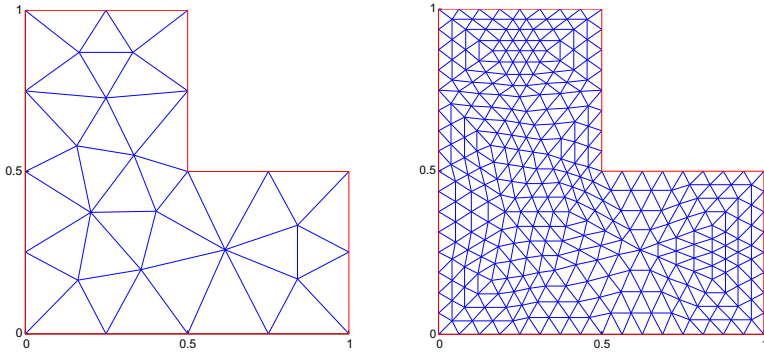


Fig. 1. Coarsest mesh (left). Triangulation obtained after two refinements (right).

Table 1. Values of γ^2

	$\nu = 0.25$	$\nu = 0.4$	$\nu = 0.49$	$\nu = 0.499$	$\nu = 0.49999$
$\ell = 0$	0.0552	0.0201	0.0019	1.8846×10^{-4}	1.8833×10^{-6}
$\ell = 1$	0.0588	0.0219	0.0021	2.0869×10^{-4}	2.0859×10^{-6}
$\ell = 2$	0.0606	0.0228	0.0022	2.1851×10^{-4}	2.1842×10^{-6}
$\ell = 3$	0.0627	0.0237	0.0023	2.2878×10^{-4}	2.2869×10^{-6}

from level $(\ell - 1)$ into four congruent triangles. The values of the constant γ and the spectral condition numbers have been computed using MATLAB.

In Table 1 the true (observed) values of γ^2 for the inequality in Lemma 2 are listed for different levels of refinement. It is evident that γ is uniformly bounded with respect to the mesh size and also with respect to the Lamé parameters (see Remark 2). The relative spectral condition number of the proposed preconditioner $\kappa(B^{-1}A) = O(1)$. The numerical values reported in Table 2 confirm the uniform bounds given in Theorem 2.

Table 2. Tabulated values of $\kappa(B^{-1}A)$

	$\nu = 0.25$	$\nu = 0.4$	$\nu = 0.49$	$\nu = 0.499$	$\nu = 0.49999$
$\ell = 0$	1.6141	1.3302	1.0910	1.0278	1.0027
$\ell = 1$	1.6405	1.3472	1.0960	1.0293	1.0029
$\ell = 2$	1.6534	1.3554	1.0983	1.0300	1.0030
$\ell = 3$	1.6683	1.3641	1.1006	1.0307	1.0030

4 Concluding Remarks

It is shown in [1] (Lemma 4.13) that the subproblem on $\mathcal{Z}$ is well conditioned and its solution can be done efficiently. Hence, the only remaining issue is to construct a uniform preconditioner for the subproblem on the space $\mathbf{V}^{\text{CR}}$.

For the case of Dirichlet boundary conditions on the entire boundary it is known how to construct optimal order multilevel preconditioners that are robust with respect to the parameter λ , (see [3] and [7]). For mixed boundary conditions or pure Neumann boundary conditions (the traction free case), however, it is much more difficult to devise a robust optimal order method. This question is subject of current research work.

Acknowledgments

The first author was supported by the Spanish MEC under projects MTM2008-03541 and HI2008-0173. The work of the fourth author has been supported in part by the US National Science Foundation, Grants DMS-0810982, and DMS-0749202. We also gratefully acknowledge the support by the Austrian Science Fund, Grants P19170-N18 and P22989-N18, and Bulgarian NSF, Grant DO 02-338/08.

References

1. Ayuso, B., Georgiev, I., Kraus, J., Zikatanov, L.: A Subspace correction method for discontinuous Galerkin discretizations of linear elasticity equations. RICAM-Report, 16-2009, Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria (2009)
2. Ayuso de Dios, B., Zikatanov, L.: Uniformly convergent iterative methods for discontinuous Galerkin discretizations. *J. Sci. Comput.* 40(1-3), 4–36 (2009)
3. Blaheta, R., Margenov, S., Neytcheva, M.: Aggregation-based multilevel preconditioning of non-conforming FEM elasticity problems. In: Dongarra, J., Madsen, K., Waśniewski, J. (eds.) PARA 2004. LNCS, vol. 3732, pp. 847–856. Springer, Heidelberg (2006)
4. Brenner, S., Scott, L.: The mathematical theory of finite element methods. Texts in Applied Mathematics, vol. 15. Springer, Heidelberg (1994)
5. Brenner, S.C., Sung, L.-Y.: Linear finite element methods for planar linear elasticity. *Math. Comp.* 59(200), 321–338 (1992)
6. Falk, R.S.: Nonconforming finite element methods for the equations of linear elasticity. *Math. Comp.* 57(196), 529–550 (1991)
7. Georgiev, I., Kraus, J.K., Margenov, S.: Multilevel preconditioning of Crouzeix-Raviart 3D pure displacement elasticity problems. In: Lirkov, I., Margenov, S., Waśniewski, J. (eds.) LSSC 2009. LNCS, vol. 5910, pp. 100–107. Springer, Heidelberg (2010)
8. Hansbo, P., Larson, M.G.: Discontinuous Galerkin and the Crouzeix-Raviart element: application to elasticity. *M2AN Math. Model. Numer. Anal.* 37(1), 63–72 (2003)
9. Kraus, J.K., Margenov, S.: Robust Algebraic Multilevel Methods and Algorithms. Radon Series on Computational and Applied Mathematics, vol. 5. Walter de Gruyter Inc., New York (October 2009)

Merger Bound States in $0 - \pi$ Josephson Structures

Todor L. Boyadjiev and Hristo T. Melemov

Plovdiv University (branch Smolyan)
hristo.melemov@yahoo.com

Abstract. The possible static distributions of magnetic flux in a $0 - \pi$ Josephson junction are described as a result of a nonlinear interaction between distributions of magnetic flux in “virtual” homogeneous and π junctions. The influence of an external magnetic field on some basic stable fluxons in a $0 - \pi$ Josephson junction as well as in the corresponding “virtual” junctions has been studied.

1 Preliminary Notes and Definitions

Note that Josephson junctions have been studied by many authors. For example, the static distributions of magnetic flux in non-homogeneous Josephson junctions are examined in [1], and the half-integer vortices in $0 - \pi$ Josephson junctions are discussed theoretically and observed experimentally in [2], [3] and [4].

According to [5], the static distributions of magnetic flux φ in homogeneous Josephson junction of length $2l$ are modeled by the following nonlinear boundary-value problem:

$$-\varphi_{xx} + \sin \varphi - \gamma = 0, \quad x \in (-l, l), \quad (1a)$$

$$\varphi_x(-l) = h_e, \quad (1b)$$

$$\varphi_x(l) = h_e, \quad (1c)$$

where h_e is the external magnetic field, γ – the external current.

Now, we consider the nonlinear equation (1a) at zero external current ($\gamma = 0$) on an infinite interval ($l \rightarrow \infty$).

The vortex distributions of the magnetic flux in the junction are solutions of equation (1a). They are very important from physical point of view. The simplest vortex is a one-fluxon (anti-fluxon) solution and we denote it by Φ_0^1 (Φ_0^{-1}) (subscript 0 denotes that the distribution is in a homogeneous junction). These solutions could be expressed in the form:

$$\Phi_0^1(x) = 4 \arctan \exp \{ (x + \xi) \}, \quad (2)$$

$$\Phi_0^{-1}(x) = 4 \arctan \exp \{ -(x + \xi) \} - 2\pi, \quad (3)$$

where ξ is a real parameter.

If we ignore some physical effects at the point of contact in barrier layers, then the static distributions of the magnetic flux in the $0 - \pi$ junction ($0 - \pi$ JJ) are modeled by the following equations [3]:

$$-\varphi_{0,xx} + \sin \varphi_0 - \gamma = 0, \quad x \in (-l, 0), \quad (4a)$$

$$-\varphi_{\pi,xx} + \sin(\varphi_\pi + \pi) - \gamma = 0, \quad x \in (0, l). \quad (4b)$$

The magnetic flux (φ_0, φ_π) and the internal magnetic field $(\varphi_{0,x}, \varphi_{\pi,x})$ at the center $x = 0$ satisfy the following continuity conditions:

$$\varphi_0(0) - \varphi_\pi(0) = 0, \quad (5a)$$

$$\varphi_{0,x}(0) - \varphi_{\pi,x}(0) = 0. \quad (5b)$$

The internal magnetic field in junctions of finite length $l < \infty$ and overlap geometry ([5]) satisfy the following boundary conditions:

$$\varphi_{0,x}(-l) = h_e, \quad (6a)$$

$$\varphi_{\pi,x}(l) = h_e, \quad (6b)$$

Equations (4)–(6) define the nonlinear boundary-value problem, corresponding to the studied model of $0 - \pi$ JJ.

Let us note that the basic numerical characteristics of an arbitrary solution of the nonlinear problem (4)–(6) are the full flux $\Delta\varphi$ and the average magnetic flux $N[\varphi]$.

In this paper we describe the static distributions of the magnetic flux in a $0 - \pi$ Josephson junction at zero external current ($\gamma = 0$) as a result of nonlinear interaction between fluxons in “virtual” homogeneous and π junctions at point $x = 0$. Mathematically, it means that for every given solution of the nonlinear boundary-value problem (4)–(6) we could find the corresponding solutions of the boundary-value problems in both homogeneous and π junctions.

We use a continuous analog of Newton method combined with spline-collocation scheme for numerical solving of the nonlinear boundary-value problem (4)–(6).

For the obtained solutions in $0 - \pi$ Josephson junction, the corresponding distributions in homogeneous and π junctions are found by solving numerically Cauchy problem with an additional condition.

We will study the influence of an external magnetic field on the basic stable fluxons in $0 - \pi$ JJ as well as in the corresponding “virtual” junctions.

2 Main Results

2.1 Statement of the Problem

We describe the static distributions of the magnetic flux in a $0 - \pi$ Josephson junction at zero external current ($\gamma = 0$) as a nonlinear interaction between distributions of the magnetic flux in “virtual” homogeneous and π junctions. For

this aim, we will give conditions which allow us to present any solution of the nonlinear boundary-value problem (4)–(6) in terms of the solutions of boundary-value problems which describe the distributions in homogeneous and π “virtual” junctions. We note that the lengths of “virtual” junctions are different from the length of a $0 - \pi$ junction.

Let the couple (φ_0, φ_π) be a solution of the nonlinear boundary problem (4)–(6) in $0 - \pi$ JJ, where $\varphi_0(x)$ satisfies the boundary condition (6a). From the continuity conditions (5) we have:

$$\varphi_0(0) = \varphi_\pi(0), \quad (7a)$$

$$\varphi_{0,x}(0) = \varphi_{\pi,x}(0). \quad (7b)$$

The function $\varphi_0(x)$ defined in the interval $(-l, 0)$, is a solution of equation (4a) with boundary conditions (6a) and (7b), and satisfies additional condition (7a). To find the solution of (1) in a homogeneous junction which participates in the construction of (φ_0, φ_π) , we are looking for the solution $\phi_0(x)$ of equation (4a), which satisfies the following conditions:

$$\phi_0(0) = \varphi_0(0), \quad \phi_{0,x}(0) = \varphi_{0,x}(0),$$

$$\phi_{0,x}(l_0) = h_e,$$

where l_0 is an unknown constant.

Analogously, the function $\varphi_\pi(x)$, defined in the interval $(0, l)$, is a solution of equation (4b) with boundary conditions (6b) and (7b) and satisfies additional condition (7a). In this case, to find the solution in π junction, we are looking for the solution $\phi_\pi(x)$ of the equation (4b) which satisfies the following conditions:

$$\phi_\pi(0) = \varphi_\pi(0), \quad \phi_{\pi,x}(0) = \varphi_{\pi,x}(0),$$

$$\phi_{\pi,x}(-l_\pi) = h_e,$$

where l_π is an unknown constant.

Finding solutions in homogeneous and π junctions is reduced to solving Stefan’s problem with an unknown right (left) boundary. There are two methods for solving:

Method 1. Method of the Cauchy problem with an additional condition.

To find function $\phi_0(x)$, we solve the Cauchy problem:

$$-\phi_{0,xx} + \sin \phi_0 = 0, \quad x \in (0, l_0), \quad (8a)$$

$$\phi_0(0) = \varphi_0(0), \quad (8b)$$

$$\phi_{0,x}(0) = \varphi_{0,x}(0), \quad (8c)$$

with the additional condition

$$\phi_{0,x}(l_0) = h_e,$$

where l_0 is an unknown constant.

The function $\phi_\pi(x)$ is a solution of the Cauchy problem:

$$-\phi_{\pi,xx} + \sin(\phi_\pi + \pi) = 0, \quad x \in (-l_\pi, 0), \quad (9a)$$

$$\phi_\pi(0) = \varphi_\pi(0), \quad (9b)$$

$$\phi_{\pi,x}(0) = \varphi_{\pi,x}(0), \quad (9c)$$

with the additional condition

$$\phi_{\pi,x}(-l_k) = h_e,$$

where l_π is an unknown constant.

Method 2. Method for solving nonlinear eigenvalue problems.

In homogeneous junction, function $\phi_0(x)$ is a solution of the following nonlinear eigenvalue problem:

$$\phi_0(0) = \varphi_0(0), \quad (10a)$$

$$\phi_{0,x}(0) = \varphi_{0,x}(0), \quad (10b)$$

$$-\phi_{0,xx} + \sin \phi_0 = 0, \quad x \in (0, l_0), \quad (10c)$$

$$\phi_{0,x}(l_0) = h_e, \quad (10d)$$

where l_0 is an unknown constant.

The function $\phi_\pi(x)$ is a solution of the following eigenvalue problem:

$$\phi_\pi(0) = \varphi_\pi(0) \quad (11a)$$

$$\phi_{\pi,x}(0) = \varphi_{\pi,x}(0), \quad (11b)$$

$$\phi_{\pi,xx} + \sin(\phi_\pi + \pi) = 0, \quad x \in (-l_\pi, 0), \quad (11c)$$

$$\phi_{\pi,x}(-l_\pi) = h_e, \quad (11d)$$

where l_π is an unknown constant.

The application of any of the above described methods gives us the functions $\phi_0(x)$ and $\phi_\pi(x)$ as well as the solutions of Stefan's problems.

So, we construct the function $\Phi_0(x)$ by means of the equalities:

$$\Phi_0(x) = \begin{cases} \varphi_0(x), & x \in (-l, 0] : \\ \phi_0(x), & x \in (0, l_0] : \end{cases}$$

which is the solution of the following boundary-value problem:

$$-\Phi_{xx} + \sin \Phi = 0, \quad x \in (-l, l_0), \quad (12a)$$

$$\Phi_x(-l) = h_e, \quad (12b)$$

$$\Phi_x(l_0) = h_e. \quad (12c)$$

In an analogous way, we obtain function $\Phi_k(x)$ with equalities:

$$\Phi_\pi(x) = \begin{cases} \phi_\pi(x) & x \in (-l_\pi, 0] : \\ \varphi_\pi(x) & x \in (0, l] : \end{cases}$$

which is the solution of the following boundary-value problem:

$$-\Phi_{xx} + \sin(\Phi + \pi) = 0, \quad x \in (-l_\pi, l), \quad (13a)$$

$$\Phi_x(-l_\pi) = h_e, \quad (13b)$$

$$\Phi_x(l) = h_e. \quad (13c)$$

The solution (φ_0, φ_π) of the nonlinear boundary-value problem (4)–(6) in $0 - \pi$ JJ is obtained from the solution $\Phi_0(x)$ of nonlinear boundary-value problem (12) and the solution $\Phi_\pi(x)$ of the nonlinear boundary-value problem (13).

The three boundary problems (4)–(6), (12) and (13) are defined in the intervals $[-l, l]$, $[-l, l_0]$ and $[-l_\pi, l]$.

2.2 Numerical Results

The solutions in homogeneous and π junctions at zero external current ($\gamma = 0$) can be obtained by the help of elliptic functions. Meanwhile, Cauchy problems with additional conditions (8) and (9) have countable sets of solutions. The solutions that determine distributions of magnetic flux in “virtual” homogeneous and π junctions depend on the value of their numerical characteristics—functionals of full energy, full magnetic flux and average magnetic flux.

We will study the basic stable distributions of the magnetic flux in $0 - \pi$ JJ of length of $2l = 16$. We denote by $S^{k,n} = \Phi^k \wedge \Phi^n$ the basic stable distributions where Φ^n and Φ^k are fluxons of “virtual” junctions.

The value of the average magnetic flux of an n -fluxon distribution Φ^n for an arbitrary h_e is a constant, i.e. $N[\Phi^n] = n$ ([6]).

The eigenvalues l_0 and l_π , defining the lengths of the “virtual” junctions, are determined in terms of the average magnetic flux

$$N_{l_0}[\varphi_{l_0}] = k, \quad (14)$$

$$N_{l_\pi}[\varphi_{l_\pi}] = n. \quad (15)$$

To investigate the influence of the external magnetic flux on the basic stable fluxons in $0 - \pi$ JJ, we have to study the behavior of the solutions for the following values of the external magnetic field:

- bifurcation points for the minimal and maximal external magnetic field;
- values of external magnetic field, at which there is a change in the number of points in which internal magnetic field is equal to external magnetic field.

We use the full magnetic flux and the values of the left bound of fluxons to obtain the values of h_e and the above mentioned points. For this purpose, we solve the following two nonlinear eigenvalue problems:

$$-\varphi_{xx} + \sin \varphi = 0, \quad x \in (-l, l), \quad (16a)$$

$$\varphi_x(\pm l) = h_e, \quad (16b)$$

$$\Delta \varphi = \varphi(l) - \varphi(-l) = \Delta_0, \quad (16c)$$

where Δ_0 is the value of the full magnetic flux, and h_e is an unknown constant, and

$$-\varphi_{xx} + \sin \varphi = 0, \quad x \in (-l, l), \quad (17a)$$

$$\varphi_x(\pm l) = h_e, \quad (17b)$$

$$\varphi(-l) = \varphi_0, \quad (17c)$$

where φ_0 is the solution at point $-l$, and h_e is an unknown constant.

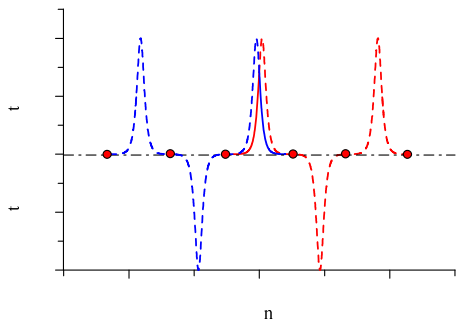


Fig. 1. Distribution of the internal magnetic field of $S^{1,1}$ for $h_e = 0$

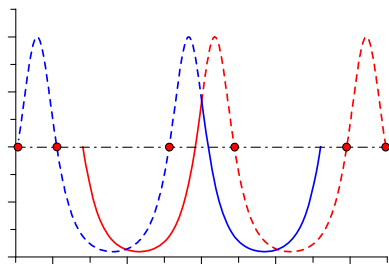


Fig. 2. Distribution of the internal magnetic field of $S^{1,1}$ for $h_e = 1$

Initially, we will describe the behavior of the fluxon $S^{1,1} = \Phi^1 \wedge \Phi^1$. If the values of the external magnetic field are nonpositive, then the first eigenvalue $l_0 \approx 9.76$ defines the length of the “virtual” homogeneous junction.

On Fig. 1, the internal magnetic field of fluxon $S^{1,1}$ in zero magnetic field $h_e = 0$ is plotted by the continuous curve (a, b). The solutions of nonlinear problems (8) and (9) for homogeneous and π junctions up to the third eigenvalues are plotted by the dashed curves c and d.

For $h_e \approx 0.001$, as a result of the minimal positive external magnetic field, the internal magnetic field of the fluxon $S^{1,1}$ has values equal to h_e in the neighbourhood of the endpoints of the junction $x = \pm l$. At this point, the value of the magnetic field in the left bound of the junction is equal to $S^{1,1}(-l) = 0$, and the full magnetic field of the solution is $\Delta S^{1,1} = 0.5$. So, we could define this value of the magnetic field and the solution of (1) at that point by solving nonlinear eigenvalue problems (16) and (17).

The second eigenvalue determines the length of the “virtual” junction for $0.001 \leq h_e \leq 1.42$. On Fig. 2, the behaviour of the external magnetic field of the fluxon $S^{1,1}$ is shown, as well as the behavior of the corresponding “virtual” junctions for $h_e = 1$. On Fig. 4, the relationship between the first (ζ_1), the second (ζ_2) and the third (ζ_3) eigenvalues of the external magnetic field is illustrated.

For the value of the external magnetic field $h_e \approx 1.42$, the value of the internal magnetic field is equal to h_e at the center of the junction $x = 0$. In this case,

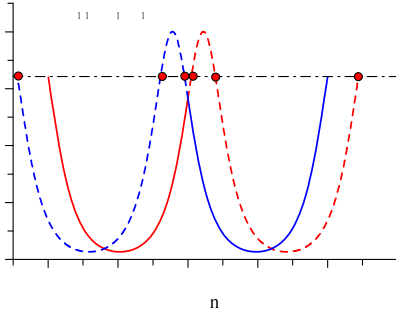


Fig. 3. Distribution of the internal magnetic field of $S^{1,1}$ for $h_e = 1.6$

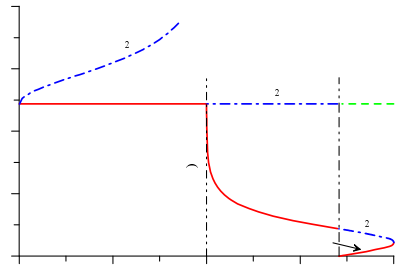


Fig. 4. Relationship between (ζ_1) and (ζ_2) of the external magnetic field for fluxon $S^{1,1}$

the value of the full magnetic flux is equal to $\Delta S^{1,1} = 1$, and the value of the magnetic flux at the left bound of the junction is equal to $S^{1,1}(-l) = \pi/2$.

For the value of the external magnetic field $h_e \geq 1.42$, the third eigenvalue determines the length of the “virtual” junction. On Fig. 3 the behavior of the fluxon $S^{1,1}$ with external magnetic field $h_e = 1.6$ is drawn.

In the case when the external field is at a neighbourhood of its bifurcation points $h_{e,cr} \approx \mp 2$, the first and the second eigenvalues approach the same value and the full magnetic fluxes at that point are $\Delta S^{1,1} = -1$ and $\Delta S^{1,1} = 3/2$, respectively, and the value at the left border of the junction is equal to $S^{1,1}(\mp l) = \mp \pi$.

The graphs of the relationships between the eigenvalues of nonlinear problem (9) for π junction are symmetric with respect to the x -axis (Fig. 4). The behaviour of the fluxon $S^{2,2} = \Phi^2 \wedge \Phi^2$ is similar.

We note that for the fluxon $S^{2,1} = \Phi^2 \wedge \Phi^1$, the graphs of the dependence of the eigenvalues on h_e in both “virtual” junctions are not symmetric with respect to the y -axis because the solutions are different. Since the two-fluxon Φ^2 has a bigger minimal external magnetic field than Φ^1 , there is bifurcation for $h_{e,cr} \approx 0.38$.

If the external magnetic field is $h_e \leq 0.4$, then the first eigenvalue (ζ_1) in the homogeneous junction and the second eigenvalue (η_2) in the π junction define the lengths of the “virtual” junctions.

For values of the external magnetic field $h_e \approx 0.4$, the value at the left bound of the junction is $S^{2,1}(-l) = 0$.

For $0.4 \leq h_e \leq 1.49$, the second eigenvalues determine the lengths of the “virtual” junctions. The dependence of the external magnetic field on the first, on the second, and on the third eigenvalues are shown on the Fig. 6. The curves ζ_1 , ζ_2 , and ζ_3 are for the homogeneous junction, the curves η_1 , η_2 , and η_3 are for the π junction.

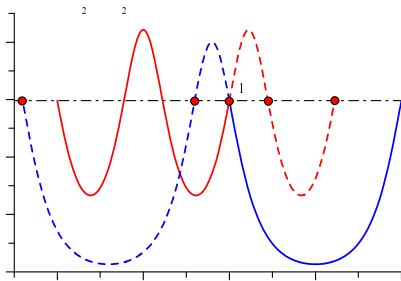


Fig. 5. Distribution of the internal magnetic field of $S^{2,1}$ for $h_e \approx 1.49$

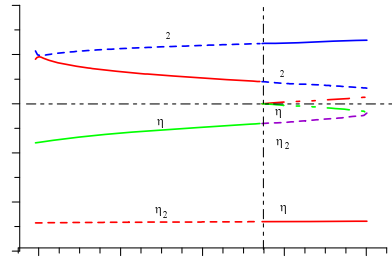


Fig. 6. Dependence of (ζ_1) (ζ_2) on the external magnetic field for fluxon $S^{2,1}$

For $h_e \approx 1.49$, the value of the internal magnetic field at point $x = 0$ is equal to the value of the external magnetic field. In this case, the first two eigenvalues are equal to zero, i.e. $\zeta_1 = \eta_1 = 0$, and the full magnetic flux is $\Delta S^{2,1} = 2$. For bigger values, the third eigenvalues determine the length of the “virtual” junctions (see on Fig. 5)

At the maximal value of the external magnetic field $h_{e,cr} \approx 2$ the first two eigenvalues are equal ($\eta_1 = \eta_2$) in the “virtual” π junction since Φ^1 in the π junction has a smaller maximal external magnetic field.

The lengths of the “virtual” homogeneous and π junctions are independent on the external magnetic field h_e for the fluxon distribution of the magnetic flux $S^{1,1}$.

Acknowledgments

Research was partially supported by Grant No. RS09FMI064, Plovdiv University, Plovdiv, Bulgaria.

References

1. Gal'pern, Yu.S., Filippov, A.T.: Bounded soliton states in inhomogeneous junctions. Sov. Phys. JETP 59 (1984) (in Russian)
2. Goldobin, E., Koelle, D., Kleiner, R.: Semifluxons in long Josephson $0-\pi$ junctions. Phys. Rev. B 66, 100508 (2002)
3. Goldobin, E., Koelle, D., Kleiner, R.: Ground states of one and two fractal vortices in long Josephson $0-k$ junctions. Phys. Rev. ser. B 70, 174519 (2004)
4. Golobdin, E., Sterk, A., Gaber, T., Koelle, D., Kleiner, R.: Dynamics of semifluxons in NG long Josephson $0-\pi$ junction. Physical Review Letters 92, 057005 (2004)
5. Licharev, K.K.: Dynamics of Josephson Junctions and Circuits. Gordon and Breach, New York (1986)
6. Atanasova, P.H., Boyadjiev, T.L., Dimova, S.N.: Numerical modeling of critical relationships of symmetric two-layer Josephson junctions. Izv. OIAI, R11-2005-162, Dubna (2005) (in Russian)

Some Error Estimates for the Discretization of Parabolic Equations on General Multidimensional Nonconforming Spatial Meshes

Abadallah Bradji¹ and Jürgen Fuhrmann²

¹ Department of Mathematics, University of Annaba–Algeria

`bradji@cmi.univ-mrs.fr`

`http://www.cmi.univ-mrs.fr/~bradji`

² Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin–Germany

`fuhrmann@wias-berlin.de`

`http://www.wias-berlin.de/~fuhrmann`

Abstract. This work is devoted to error estimates for the discretization of parabolic equations on general nonconforming spatial meshes in several space dimensions. These meshes have been recently used to approximate stationary anisotropic heterogeneous diffusion equations and nonlinear equations. We present an implicit time discretization scheme based on an orthogonal projection of the exact initial value. We prove that, when the discrete flux is calculated using a stabilized discrete gradient, the convergence order is $h_{\mathcal{D}} + k$, where $h_{\mathcal{D}}$ (resp. k) is the mesh size of the spatial (resp. time) discretization. This estimate is valid for discrete norms $\mathbb{L}^{\infty}(0, T; H_0^1(\Omega))$ and $\mathcal{W}^{1,\infty}(0, T; L^2(\Omega))$ under the regularity assumption $u \in \mathcal{C}^2([0, T]; \mathcal{C}^2(\overline{\Omega}))$ for the exact solution u . These error estimates are useful because they allow to obtain approximations to the exact solution and its first derivatives of order $h_{\mathcal{D}} + k$.

Keywords: non-conforming grid, parabolic equation, SUSHI scheme, implicit scheme, discrete gradient.

1 Introduction and Aim of This Paper

The finite volume method is well established to approximate various types of conservation laws used in many engineering fields, such as fluid mechanics, heat and mass transfer or petroleum engineering. It can be applied in arbitrary geometries and is *locally conservative*, see [6] and the references therein. In order to yield a finite volume discretization, we integrate the equation to be solved on the so called *control volumes*. We use then *numerical fluxes* to approximate, using the discrete unknowns, the continuous fluxes over the boundaries of the control volumes, which appear after the integration by parts. A widely used definition of admissible finite volume meshes for viscous conservation laws can be found in

[6, Definition 9.1, Pages 762–763]. Among the features of this definition is that control volumes are open polygonal convex sets. In addition to this, for admissibility, the mesh should satisfy an orthogonality condition, that is, there exists a family of points $(x_K)_{K \in \mathcal{T}}$, such that for a given edge σ_{KL} , the line segment $x_K x_L$ is orthogonal to this edge. This condition is useful for approximating the fluxes over a given edge using two point difference quotients. The construction of such meshes in general geometries is possible in many interesting cases but still linked to a number of challenging problems [4]. Therefore in many cases it is useful to drop the orthogonality condition and to assume general polyhedral control volumes, where the boundary of each control volume is a finite union of subsets of hyperplanes, cf. [7] and the references therein. Recently [7], a large class of nonconforming meshes could be used in the approximation of stationary anisotropic heterogeneous diffusion equations, and some error estimates have been provided.

The aim of the present paper is to consider a generalization of this approach to the nonstationary case. We regard a nonstationary heat equation and derive error estimates in $\mathbb{L}^\infty(0, T; H_0^1(\Omega))$ and $\mathcal{W}^{1,\infty}(0, T; L^2(\Omega))$. These error estimates are useful since they allow us to obtain estimates for approximations for not only the unknown solution but also its first derivatives. The present work is as well an extension of the previous papers [1,2], where we dealt with the error estimate of the finite volume approximation of parabolic equations in two or three dimensions and using admissible meshes as described in [6]. The main result of the present work is Theorem 2 below. Because of the limited number of pages, we only give a sketch of the proof. A detailed proof as well as a general framework (in which the composite scheme (14)–(15) is a particular case) of the present work is the subject of the paper under preparation [3].

2 Equation to be Solved and Preliminaries

The present work deals with following multidimensional transient diffusion problem:

$$u_t(x, t) - \Delta u(x, t) = f(x, t), \quad (x, t) \in \Omega \times (0, T), \quad (1)$$

where, Ω is an open bounded polyhedral subset in $\mathbb{R}^d$, with $d \in \mathbb{N}^*$, $T > 0$, and f is a given function.

An initial condition is given by:

$$u(x, 0) = u^0(x), \quad x \in \Omega. \quad (2)$$

A Dirichlet boundary condition is defined by

$$u(x, t) = 0, \quad (x, t) \in \partial\Omega \times (0, T), \quad (3)$$

where, we denote by $\partial\Omega = \overline{\Omega} \setminus \Omega$ the boundary of Ω .

To define a weak solution for (1)–(3) and throughout our work, we use the standard notation for function space: $\mathbb{L}^p(\omega)$ –Lebesgue space, $\mathcal{W}^{k,p}(\omega)$,

$H^k(\omega) = \mathcal{W}^{2,p}(\omega)$ Sobolev spaces, $\mathbb{L}^p(0, T; X)$ –Bochner space of functions defined in $(0, T)$ with values in a Banach space X , where k integer and $p \in [1, +\infty]$.

For $p \in [1, +\infty)$, the space $\mathbb{L}^p(0, T; X)$ is equipped with the norm:

$$\|u\|_{\mathbb{L}^p(0,T;X)} = \left(\int_0^T \|u\|_X^p dt \right)^{\frac{1}{p}}. \quad (4)$$

We define the space $\mathcal{W}^{m,p}(0, T; X)$ as:

$$\mathcal{W}^{m,p}(0, T; X) = \left\{ u \in \mathbb{L}^p(0, T; X); \frac{d^j u}{dt^j} \in \mathbb{L}^p(0, T; X), j \in \llbracket 1, m \rrbracket \right\}. \quad (5)$$

The space $\mathcal{W}^{m,p}(0, T; X)$ is equipped with the norm:

$$\|u\|_{\mathcal{W}^{m,p}(0,T;X)} = \left(\sum_{j=1}^m \left\| \frac{d^j u}{dt^j} \right\|_{\mathbb{L}^p(0,T;X)}^p \right)^{\frac{1}{p}}. \quad (6)$$

The spaces $\mathbb{L}^\infty(0, T; X)$ and $\mathcal{W}^{m,\infty}(0, T; X)$ can be defined in a similar way, see for instance [8, Pages 47–48] and [5, Pages 285–286].

The following Theorem gives a sense for a weak solution for problem (1)–(3) (recall that $H^{-1}(\Omega)$ is the dual of $H_0^1(\Omega)$):

Theorem 1. (cf. [5, Theorems 3 and 4, Pages 356–358]) Let $f \in \mathbb{L}^2(0, T; \mathbb{L}^2(\Omega))$ and $u^0 \in \mathbb{L}^2(\Omega)$. Then, there exists a unique weak solution for (1)–(3) in the following sense: there exists a function $u \in \mathbb{L}^2(0, T; H_0^1(\Omega))$ such that $u_t \in \mathbb{L}^2(0, T; H^{-1}(\Omega))$ and:

(i) For a.e. $0 \leq t \leq T$

$$\langle u_t, v \rangle + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx, \text{ for } \forall v \in H_0^1(\Omega) \quad (7)$$

(ii)

$$u(0) = u^0. \quad (8)$$

Remark 1. According to assumptions of Theorem 1, since $u \in \mathbb{L}^2(0, T; H_0^1(\Omega))$ and $u_t \in \mathbb{L}^2(0, T; H^{-1}(\Omega))$, one could obtain (cf. [5, Theorems 3, Pages 287]) $u \in \mathcal{C}([0, T]; \mathbb{L}^2(\Omega))$, and thus equation (8) makes sense.

The convergence of the finite volume scheme we want to present is analyzed using the space $\mathcal{C}^m([0, T]; X)$ of m –times continuously differentiable mappings of the interval $[0, T]$ with values in X .

The space $\mathcal{C}^m([0, T]; X)$ is equipped with the norm

$$\|u\|_{\mathcal{C}^m([0,T];X)} = \max_{j \in \llbracket 1, m \rrbracket} \left\{ \sup_{t \in [0, T]} \left\| \frac{d^j u}{dt^j}(t) \right\|_X \right\}, \quad (9)$$

where $\|\cdot\|_X$ denotes the norm of X .

Throughout the convergence analysis of the finite volume scheme, the space X is often a space of the form $\mathcal{C}^m(\overline{\Omega})$ where $m \in \mathbb{N}$.

3 Meshes and Schemes

Definition 1 (Space discretization, cf. [7]). Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. A discretisation of Ω , denoted by $\mathcal{D}$, is defined as the triplet $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where:

1. $\mathcal{M}$ is a finite family of non empty connected open disjoint subsets of Ω (the “control volumes”) such that $\overline{\Omega} = \cup_{K \in \mathcal{M}} \overline{K}$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K ; let $m(K) > 0$ denote the measure of K and h_K denote the diameter of K .
2. $\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the “edges” of the mesh), such that, for all $\sigma \in \mathcal{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$ -dimensional measure is strictly positive. We also assume that, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$. For any $\sigma \in \mathcal{E}$, we denote by $\mathcal{M}_\sigma = \{K; \sigma \in \mathcal{E}_K\}$. We then assume that, for any $\sigma \in \mathcal{E}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \subset \partial\Omega$ (the set of these interfaces, called boundary interfaces, denoted by $\mathcal{E}_{\text{ext}}$) or $\mathcal{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, denoted by $\mathcal{E}_{\text{int}}$). For all $\sigma \in \mathcal{E}$, we denote by x_σ the barycentre of σ . For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K .
3. $\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (x_K)_{K \in \mathcal{M}}$, such that for all $K \in \mathcal{M}$, $x_K \in K$ and K is assumed to be x_K -star-shaped, which means that for all $x \in K$, the property $[x_K, x] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between x_K and the hyperplane including σ , one assumes that $d_{K,\sigma} > 0$. We then denote by $\mathcal{D}_{K,\sigma}$ the cone with vertex x_K and basis σ .

The discretization of Ω is then performed using the mesh $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ described in Definition 1, whereas the time discretization is performed with a constant time step $k = \frac{T}{N+1}$, where $N \in \mathbb{N}^*$, and we shall denote $t_n = nk$, for $n \in \llbracket 0, N+1 \rrbracket$. Throughout this paper, the letter C stands for a positive constant independent of the parameters of the space and time discretizations.

Let $\mathcal{X}_{\mathcal{D}}$ be the set of all $((v_K)_{K \in \mathcal{M}}, (v_\sigma)_{\sigma \in \mathcal{E}})$, and let $\mathcal{X}_{\mathcal{D},0} \subset \mathcal{X}_{\mathcal{D}}$ be the set of all $v \in \mathcal{X}_{\mathcal{D}}$ such that $v_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\text{ext}}$. The space $\mathcal{X}_{\mathcal{D}}$ is equipped with the semi-norm $|v|_{\mathcal{X}}^2 = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} \frac{m(\sigma)}{d_{K,\sigma}} (v_\sigma - v_K)^2$. For a given family of real numbers $\{\beta_\sigma^K; K \in \mathcal{M}, \sigma \in \mathcal{E}_{\text{int}}\}$, with $\beta_\sigma^K \neq 0$ only for some control volumes which are “close” to σ , and such that

$$1 = \sum_{K \in \mathcal{M}} \beta_\sigma^K \quad \text{and} \quad x_\sigma = \sum_{K \in \mathcal{M}} \beta_\sigma^K x_K, \quad (10)$$

we define a space with dimension smaller than that of $\mathcal{X}_{\mathcal{D},0}$. This can be achieved by expressing u_σ , for all $\sigma \in \mathcal{B}$, where $\mathcal{B} \subset \mathcal{E}_{\text{int}}$ as a consistent barycentric combination of the values u_K , i.e., $u_\sigma = \sum_{K \in \mathcal{M}} \beta_\sigma^K u_K$.

We decompose then the set $\mathcal{E}_{\text{int}}$ of interfaces into two non intersecting subsets, that is: $\mathcal{E}_{\text{int}} = \mathcal{B} \cup \mathcal{H}$ and $\mathcal{H} = \mathcal{E}_{\text{int}} \setminus \mathcal{B}$. The interface unknowns associated with $\mathcal{B}$ will be computed by using the barycentric formula $u_\sigma = \sum_{K \in \mathcal{M}} \beta_\sigma^K u_K$. The unknowns of the scheme will be then the quantities u_K for $K \in \mathcal{M}$ and u_σ for $\sigma \in \mathcal{H}$. Consider then the space $\mathcal{X}_{\mathcal{D}, \mathcal{B}} \subset \mathcal{X}_{\mathcal{D}, 0}$ given by

$$\mathcal{X}_{\mathcal{D}, \mathcal{B}} = \left\{ v \in \mathcal{X}_{\mathcal{D}, 0} \text{ such that } v_\sigma = \sum_{K \in \mathcal{M}} \beta_\sigma^K v_K, \forall \sigma \in \mathcal{B} \right\}.$$

We define the subspace $H_{\mathcal{M}}(\Omega)$ of $\mathbb{L}^2(\Omega)$ as the set of the functions which are constant on each control volume $K \in \mathcal{M}$. We then denote, for all $v \in H_{\mathcal{M}}(\Omega)$ and for all $\sigma \in \mathcal{E}_{\text{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$, $D_\sigma v = |v_K - v_L|$ and $d_\sigma = d_{K, \sigma} + d_{L, \sigma}$, and for all $\sigma \in \mathcal{E}_{\text{ext}}$ with $\mathcal{M}_\sigma = \{K\}$, we denote $D_\sigma v = |v_K|$ and $d_\sigma = d_{K, \sigma}$. We then define the following norm:

$$\forall v \in H_{\mathcal{M}}(\Omega), \|v\|_{1,2,\mathcal{M}}^2 = \sum_{\sigma \in \mathcal{E}} m(\sigma) \frac{(D_\sigma v)^2}{d_\sigma}. \quad (11)$$

For all $v \in \mathcal{X}_{\mathcal{D}}$, we denote by $\Pi_{\mathcal{M}} v \in H_{\mathcal{M}}(\Omega)$ the piecewise constant function from Ω to $\mathbb{R}$ defined by $\Pi_{\mathcal{M}} v(x) = v_K$, for a.e. $x \in K$, for all $K \in \mathcal{M}$. For all $\varphi \in \mathcal{C}(\Omega)$, we denote by $\mathcal{P}_{\mathcal{D}} \varphi$ the element of $\mathcal{X}_{\mathcal{D}}$ defined by $((\varphi(x_K))_{K \in \mathcal{M}}, (\varphi(x_\sigma))_{\sigma \in \mathcal{E}})$, and by $\mathcal{P}_{\mathcal{D}, \mathcal{B}} \varphi$ the element $v \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}$ such that $v_K = \varphi(x_K)$ for all $K \in \mathcal{M}$, $v_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\text{ext}}$, $v_\sigma = \sum_{K \in \mathcal{M}} \beta_\sigma^K \varphi(x_K)$ for all $\sigma \in \mathcal{B}$ and $v_\sigma = \varphi(x_\sigma)$ for all $\sigma \in \mathcal{E}_{\text{int}} \setminus \mathcal{B}$.

We denote by $\mathcal{P}_{\mathcal{M}} \varphi \in H_{\mathcal{M}}(\Omega)$ the element defined by $\mathcal{P}_{\mathcal{M}} \varphi(x) = \varphi(x_K)$, for a.e. $x \in K$, for all $K \in \mathcal{M}$.

We need, to analyse the convergence, to consider the size of the discretization $\mathcal{D}$ defined by $h_{\mathcal{D}} = \sup\{\text{diam}(K), K \in \mathcal{M}\}$ and the regularity of the mesh is given by $\theta_{\mathcal{D}} = \max \left(\max_{\sigma \in \mathcal{E}_{\text{int}}, K, L \in \mathcal{M}} \frac{d_{K, \sigma}}{d_{L, \sigma}}, \max_{K \in \mathcal{M}, \sigma \in \mathcal{E}_K} \frac{h_K}{d_{K, \sigma}} \right)$.

For a given set $\mathcal{B} \subset \mathcal{E}_{\text{int}}$ and for a given family $(\beta_\sigma^K)_{K \in \mathcal{M}, \sigma \in \mathcal{E}_{\text{int}}}$ satisfying property (10), we introduce $\theta_{\mathcal{D}, \mathcal{B}} = \max \left(\theta_{\mathcal{D}}, \max_{K \in \mathcal{M}, \sigma \in \mathcal{E}_K \cap \mathcal{B}} \frac{\sum_{L \in \mathcal{M}} |\beta_\sigma^L| |x_\sigma - x_L|^2}{h_K^2} \right)$.

The scheme we want to consider in this note (A general framework will be detailed in a future paper.) is based on the use of the discrete gradient given in [7]. For $u \in \mathcal{X}_{\mathcal{D}}$, we define, for all $K \in \mathcal{M}$

$$\nabla_{\mathcal{D}} u(x) = \nabla_{K, \sigma} u, \text{ a. e. } x \in \mathcal{D}_{K, \sigma}, \quad (12)$$

where $\mathcal{D}_{K, \sigma}$ is the cone with vertex x_K and basis σ and

$$\nabla_{K, \sigma} u = \nabla_K u + \left(\frac{\sqrt{d}}{d_{K, \sigma}} (u_\sigma - u_K - \nabla_K u \cdot (x_\sigma - x_K)) \right) \mathbf{n}_{K, \sigma}, \quad (13)$$

where $\nabla_K u = \frac{1}{m(K)} \sum_{\sigma \in \mathcal{E}_K} m(\sigma) (u_\sigma - u_K) \mathbf{n}_{K, \sigma}$ and d is the space dimension.

Using these notations, we can consider now the discrete problem as follows: For any $n \in \llbracket 0, N \rrbracket$, find $u_{\mathcal{D}}^n \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}$ such that

$$\begin{aligned} & (\partial^1 \Pi_{\mathcal{M}} u_{\mathcal{D}}^{n+1}, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)} + (\nabla_{\mathcal{D}} u_{\mathcal{D}}^{n+1}, \nabla_{\mathcal{D}} v)_{(\mathbb{L}^2(\Omega))^d} \\ & = \sum_{K \in \mathcal{M}} m(K) f_K^n v_K, \quad \forall v \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}, \end{aligned} \quad (14)$$

and find $u_{\mathcal{D}}^0 \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}$ such that

$$(\nabla_{\mathcal{D}} u_{\mathcal{D}}^0, \nabla_{\mathcal{D}} v)_{(\mathbb{L}^2(\Omega))^d} = -(\Delta u^0, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}, \quad (15)$$

where $\partial^1 v^n = \frac{v^n - v^{n-1}}{k}$, $f_K^n = \frac{1}{km(K)} \int_{t_n}^{t_{n+1}} \int_K f(x, t) dx dt$, and $(\cdot, \cdot)_{\mathbb{L}^2(\Omega)}$ (resp. $(\cdot, \cdot)_{(\mathbb{L}^2(\Omega))^d}$) denotes the $\mathbb{L}^2$ (resp. $(\mathbb{L}^2(\Omega))^d$) inner product.

4 Convergence Results

The main result of this paper is the following theorem:

Theorem 2. (Error estimates for the composite scheme (14)–(15)) *Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Assume that the weak solution of (1)–(3) in the sense of Theorem 1 satisfies $u \in \mathcal{C}^2([0, T]; \mathcal{C}^2(\overline{\Omega}))$. Let $k = \frac{T}{N+1}$, with $N \in \mathbb{N}^*$, and denote by $t_n = nk$, for $n \in \llbracket 0, N+1 \rrbracket$. Let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be a discretization in the sense of Definition 1. Let $\mathcal{B} \subset \mathcal{E}_{\text{int}}$ be given and let $\{\beta_{\sigma}^K, \sigma \in \mathcal{B}, K \in \mathcal{M}\}$ be a subset of $\mathbb{R}$ satisfying (10). Assume that $\theta_{\mathcal{D}, \mathcal{B}}$ satisfies $\theta \geq \theta_{\mathcal{D}, \mathcal{B}}$. Then there exists a unique solution $(u_{\mathcal{D}}^n)_{n \in \llbracket 0, N+1 \rrbracket}$ for (14)–(15). For each $n \in \llbracket 0, N+1 \rrbracket$, let us define the error $e_{\mathcal{M}}^n \in H_{\mathcal{M}}(\Omega)$ by:*

$$e_{\mathcal{M}}^n = \mathcal{P}_{\mathcal{M}} u(\cdot, t_n) - \Pi_{\mathcal{M}} u_{\mathcal{D}}^n. \quad (16)$$

Then, the following error estimates hold

– discrete $\mathbb{L}^\infty(0, T; H_0^1(\Omega))$ -estimate: for all $n \in \llbracket 0, N+1 \rrbracket$

$$\|e_{\mathcal{M}}^n\|_{1,2,\mathcal{M}} \leq C(h_{\mathcal{D}} + k) \|u\|_{\mathcal{C}^2([0,T]; \mathcal{C}^2(\overline{\Omega}))}. \quad (17)$$

– $\mathcal{W}^{1,\infty}(0, T; \mathbb{L}^2(\Omega))$ -estimate: for all $n \in \llbracket 1, N+1 \rrbracket$

$$\|\partial^1 e_{\mathcal{M}}^n\|_{\mathbb{L}^2(\Omega)} \leq C(h_{\mathcal{D}} + k) \|u\|_{\mathcal{C}^2([0,T]; \mathcal{C}^2(\overline{\Omega}))}, \quad (18)$$

where $\partial^1 v^n = \frac{v^n - v^{n-1}}{k}$.

– error estimate in the gradient approximation: for all $n \in \llbracket 0, N+1 \rrbracket$

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}}^n - \nabla u(\cdot, t_n)\|_{(\mathbb{L}^2(\Omega))^d} \leq C(h_{\mathcal{D}} + k) \|u\|_{\mathcal{C}^2([0,T]; \mathcal{C}^2(\overline{\Omega}))}. \quad (19)$$

Sketch of the proof: The uniqueness of $(u_{\mathcal{D}}^n)_{n \in \llbracket 0, N+1 \rrbracket}$ satisfying (14)–(15) could be deduced from the stability [7, 37, Lemma 4.1]. As usual, we can use this uniqueness to prove the existence.

To prove (17)–(19), we compare the solution $(u_{\mathcal{D}}^n)_{n \in \llbracket 0, N+1 \rrbracket}$ satisfying (14)–(15) with the solution defined by: for any $n \in \llbracket 0, N+1 \rrbracket$, find $\bar{u}_{\mathcal{D}}^n \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}$ such that

$$(\nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}}^n, \nabla_{\mathcal{D}} v)_{(\mathbb{L}^2(\Omega))^d} = - \sum_{K \in \mathcal{M}} v_K \int_K \Delta u(x, t_n) dx, \quad \forall v \in \mathcal{X}_{\mathcal{D}, \mathcal{B}}. \quad (20)$$

Step 1. (Comparison between u and $\bar{u}_{\mathcal{D}}^n$). We use mainly the results of [7, Theorem 4.2], with some attention to be paid for the constants appear in [7, Theorem 4.2] and its related estimates, and [2] to get the following estimates

$$\|\mathcal{P}_{\mathcal{M}} u(\cdot, t_n) - \Pi_{\mathcal{M}} \bar{u}_{\mathcal{D}}^n\|_{1,2,\mathcal{M}} \leq C h_{\mathcal{D}} \|u\|_{C([0,T];C^2(\bar{\Omega}))}, \quad (21)$$

$$\|\partial^j (\mathcal{P}_{\mathcal{M}} u(\cdot, t_n) - \Pi_{\mathcal{M}} \bar{u}_{\mathcal{D}}^n)\|_{\mathbb{L}^2(\Omega)} \leq C h_{\mathcal{D}} \|u\|_{C^j([0,T];C^2(\bar{\Omega}))}, \quad j \in \llbracket 0, 2 \rrbracket, \quad (22)$$

where we have denoted $\partial^0 v^n = v^n$ and $\partial^2 v^n = \frac{1}{k} (\partial^1 v^n - \partial^1 v^{n-1})$, and

$$\|\nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}}^n - \nabla u(\cdot, t_n)\|_{(\mathbb{L}^2(\Omega))^d} \leq C h_{\mathcal{D}} \|u\|_{C^2([0,T];C^2(\bar{\Omega}))}. \quad (23)$$

Step 2. (Comparison between $\bar{u}_{\mathcal{D}}^n$ and $u_{\mathcal{D}}^n$) Using similar techniques to that of [2, (16)–(31), Pages 236–238], with some attention to be paid for the constants, and estimate (22) leads to, for all $n \in \llbracket 0, N \rrbracket$

$$\|\partial^1 \Pi_{\mathcal{M}} \eta_{\mathcal{D}}^{n+1}\|_{\mathbb{L}^2(\Omega)} \leq C(h_{\mathcal{D}} + k) \|u\|_{C^2([0,T];C^2(\bar{\Omega}))}, \quad (24)$$

where $\eta_{\mathcal{D}}^n = \bar{u}_{\mathcal{D}}^n - u_{\mathcal{D}}^n$.

This with (22) yields (18). Using similar techniques to that of [2, (33)–(35), Page 239] and [7, (75), Lemma 5.3] when $p = 2$, (24), (22), and the fact that $\eta_{\mathcal{D}}^0 = 0$ implies that, for all $n \in \llbracket 0, N+1 \rrbracket$

$$\|\Pi_{\mathcal{M}} \eta_{\mathcal{D}}^n\|_{\mathcal{X}} \leq C(h_{\mathcal{D}} + k) \|u\|_{C^2([0,T];C^2(\bar{\Omega}))}, \quad (25)$$

this with [7, (36)] yields that, for all $n \in \llbracket 0, N+1 \rrbracket$

$$\|\Pi_{\mathcal{M}} \eta_{\mathcal{D}}^n\|_{1,2,\mathcal{M}} \leq C(h_{\mathcal{D}} + k) \|u\|_{C^2([0,T];C^2(\bar{\Omega}))}. \quad (26)$$

This with (21) leads to (17). Gathering estimates (25) and [7, (37), Lemma 4.1] yields $\|\nabla_{\mathcal{D}} \eta_{\mathcal{D}}^n\|_{(\mathbb{L}^2(\Omega))^d} \leq C(h_{\mathcal{D}} + k) \|u\|_{C^2([0,T];C^2(\bar{\Omega}))}$, and then we combine this with (23) to get (19). $\square$

5 Conclusion

Because of the limited number of pages, we only considered a simple non stationary heat equation discretized on general multidimensionnal nonconforming meshes using the finite volume method based on the discrete stabilized gradient (12)–(13). More general investigations are subject of the paper under prepatation [3].

References

1. Bradji, A.: Some simples error estimates for finite volume approximation of parabolic equations. *Comptes Rendus de l'Académie de Sciences, Paris* 346(9-10), 571–574 (2008)
2. Bradji, A., Fuhrmann, J.: Some error estimates in finite volume method for parabolic equations. In: Eymard, R., Hérard, J.-M. (eds.) *Finite Volumes for Complex Applications V, Proceedings of the 5th International Symposium on Finite Volume for Complex Applications*, pp. 233–240. Wiley, Chichester (2008)
3. Bradji, A., Fuhrmann, J.: Error estimates for fully and semi-discretization schemes on general nonconforming meshes of linear parabolic equations (in progress)
4. Si, H., Gärtner, K., Fuhrmann, J.: Boundary conforming Delaunay mesh generation. *Comput. Math. Math. Phys.* 50, 38–53 (2010)
5. Evans, L.C.: *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (1998)
6. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. *Handbook of Numerical Analysis*. In: Ciarlet, P.G., Lions, J.L. (eds.), vol. VII, pp. 723–1020 (2000)
7. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. *IMA J. Numer. Anal.* (Advance Access published on June 16, 2009), doi:10.1093/imanum/drn084
8. Feistauer, M., Felcman, J., Straskraba, I.: *Mathematical and Computational Methods for Compressible Flow*. Oxford Science Publications, Oxford (2004)

Finite-Volume Difference Scheme for the Black-Scholes Equation in Stochastic Volatility Models

Tatiana Chernogorova and Radoslav Valkov

Sofia University, Faculty of Mathematics and Informatics
{chernogorova,rvalkov}@fmi.uni-sofia.bg

Abstract. We study numerically the two-dimensional Black-Scholes equation in stochastic volatility models [3]. For these models, starting from the conservative form of the equation, we construct a finite-volume difference scheme using the appropriate boundary conditions. The scheme is first order accurate in the space grid size. We also present some results from numerical experiments that confirm this.

Keywords: Black-Scholes equation, dynamical boundary condition, finite difference, finite-volume.

1 Introduction

In financial modelling, the Black-Scholes model [3, 6] for determination of the fair value of a call option or derivative security of the market has become very popular. For the Black-Scholes equation, the boundary condition is of Dirichlet type, which corresponds to the underlying asset being absorbed. However, in many situations outside the standard Black-Scholes setting, the pricing equation has degenerate, or too fast growing coefficients and standard PDE theory does not apply [7]. A such example are the Heston model [4], the CEV-model, the CIR-model, etc., see the discussion in [3]. The knowledge of the boundary behaviour is crucial when using numerical methods to calculate option prices even if these conditions are redundant from a strict mathematical point of view. Indeed, in [2, 3], boundary conditions for several pricing PDEs are discussed. The purpose of the present paper is to study numerically the PDE from [3] following the results concerning the boundary behaviour of the solution (the price) for vanishing values of the volatility. The present problem suffers from the following additional difficulty in comparison with those in [5]. The Dirichlet problem is considered and the solution space domain of the problem in [5] is the rectangle $\Omega \equiv (0, X) \times (\xi, Y)$, where $0 < X < \infty$, $0 < \xi < Y < \infty$. For our problem $\Omega \equiv (0, \infty)^2$, following [3] we consider *dynamical boundary condition at $y = 0$* .

The paper is organized as follows. In the next section we formulate the continuous problem and rewrite the differential equation in divergence form. Then, in Section 3, we derive a finite volume difference scheme based on the fitting technique of S. Wang [5, 9]. In Section 4 we perform full discretization. Numerical experiments are discussed in Section 5.

2 The Continuous Problem

We consider the Black-Scholes equation in stochastic volatility models and for clearness we take the CIR model [3] with initial and appropriate boundary conditions:

$$\frac{\partial u}{\partial t} = \frac{1}{2}x^2y \frac{\partial^2 u}{\partial x^2} + \rho x \sigma(y) \sqrt{y} \frac{\partial^2 u}{\partial x \partial y} + \frac{\sigma^2(y)}{2} \frac{\partial^2 u}{\partial y^2} + \beta(y) \frac{\partial u}{\partial y}, \quad (x, y, t) \in Q_T, \quad (1)$$

$$u(x, y, 0) = g(x), \quad (x, y) \in [0, \infty) \times [0, \infty) \equiv \Omega, \quad Q_T = \Omega \times (0, T], \quad (2)$$

$$u(0, y, t) = g(0), \quad (y, t) \in [0, \infty) \times [0, T], \quad (3)$$

$$\frac{\partial u}{\partial t}(x, 0, t) = \beta(0) \frac{\partial u}{\partial y}(x, 0, t), \quad (x, t) \in (0, \infty) \times (0, T]. \quad (4)$$

Hypothesis. The drift $\beta \in C^1([0, \infty))$ with a Holder (α) continuous derivative for some α , and $\beta(0) \geq 0$. The volatility $\sigma : [0, \infty) \rightarrow [0, \infty)$ satisfies $\sigma(0) = 0$ and $\sigma(y) > 0$ for all $y > 0$, and the function $\sigma^2(y)$ is continuously differentiable on $[0, \infty)$ with a Holder (α) continuous derivative. The growth condition $|\beta(y) + \sigma(y)| \leq C(1+y)$ holds for all $y \geq 0$ where C is a constant. The pay-off function g is bounded and it is twice continuously differentiable on $[0, \infty)$. Moreover, $xg'(x)$ and $x^2g''(x)$ are bounded.

Then, it was proved in [3] that the problem (1)-(3) has unique solution (option price) $u \in C^{2,2,1}((0, \infty)^2 \times [0, T]) \cap C^{0,1,1}((0, \infty) \times [0, \infty) \times [0, T])$, the function $0.5x^2u_{xx}$ is bounded and σ^2u_{yy} , $\sigma(y)\sqrt{y}u_{xy} \rightarrow 0$ as $y \rightarrow 0$ for any $t_0 \in [0, T]$ and any positive x_0 . Consequently, it follows that $\lim_{(x,y,t) \rightarrow (x_0,0,t_0)} (u_t(x, y, t) - \beta(y)u_y(x, y, t)) = 0$.

To assist the formulation of the finite volume method, it is convenient to write (1) in the following divergent form:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla \cdot (\mathbf{k}(u)) - pu, \quad \mathbf{k}(u) = (\mathbf{A} \nabla u + \mathbf{b}u), \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \\ a_{11} &= 0.5x^2y, \quad a_{22} = 0.5\sigma^2(y), \quad a_{12} = a_{21} = 0.5\rho x \sigma(y) \sqrt{y}, \\ \mathbf{b} &= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_1(x, y) \\ b_2(y) \end{pmatrix} = \begin{pmatrix} -xy - 0.5\rho x \sigma'(y) \sqrt{y} - 0.25\rho x \sigma(y) \frac{1}{\sqrt{y}} \\ \beta(y) - 0.5\rho \sigma(y) \sqrt{y} - \sigma(y) \sigma'(y) \end{pmatrix}, \\ p(y) &= -y - \rho \sigma'(y) \sqrt{y} - 0.5\rho \sigma(y) \frac{1}{\sqrt{y}} + \beta'(y) - \sigma'^2(y) - \sigma(y) \sigma''(y). \end{aligned}$$

3 Space Discretization

We introduce the uniform mesh $\overline{w} = \overline{w}_x \times \overline{w}_y$, $\overline{w}_x = \{x_i = ih_x, i = 0, 1, \dots, N_x, N_x h_x = X\}$, $\overline{w}_y = \{y_j = jh_y, j = 0, 1, \dots, N_y, N_y h_y = Y\}$ and the secondary mesh $x_{i\pm 1/2} = 0.5(x_{i+1} + x_i)$, $y_{j\pm 1/2} = 0.5(y_{j+1} + y_j)$, $x_{-1/2} = x_0 = 0$, $x_{N_x+1/2} = x_{N_x} = X$ [8]. For *computational purpose*, we truncate the asset regions into $[0, X]$ and $[0, Y]$. Following [2] for sufficiently large X, Y we will take

$$u(X, y, t) = g(X), u(x, Y, t) = g(x), y \in [0, Y], x \in [0, X]. \quad (5)$$

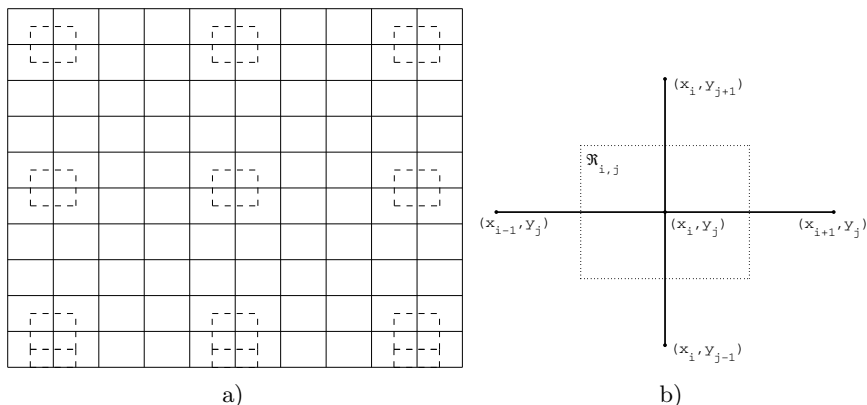


Fig. 1. Typical local structure cases of the meshes

We integrate (1) on the cell $\mathfrak{R}_{i,j} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$, $i = 1, 2, \dots, N_x - 1$; $j = 1, 2, \dots, N_y - 1$, see Fig. 1, and applying the mid point quadrature rule to the first and third terms, we obtain

$$\frac{\partial u_{i,j}}{\partial t} R_{i,j} = \int_{\mathfrak{R}_{i,j}} \nabla \cdot (\mathbf{k}(u)) dx dy - R_{i,j} p_{i,j} u_{i,j}, \quad R_{i,j} = h_x h_y.$$

Further we concentrate on the approximation of the middle term:

$$\begin{aligned} \int_{\mathfrak{R}_{i,j}} \nabla \cdot (\mathbf{k}(u)) dx dy &= \int_{\partial \mathfrak{R}_{i,j}} \mathbf{k} \cdot \mathbf{n} ds = \int_{(x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}})}^{(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})} \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u \right) dy \\ &- \int_{(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}})}^{(x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}})} \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u \right) dy + \int_{(x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}})}^{(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})} \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} + b_2 u \right) dx \\ &- \int_{(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}})}^{(x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}})} \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} + b_2 u \right) dx = I_1 - I_2 + I_3 - I_4. \end{aligned}$$

For the first integral we have: $I_1 \approx f_1|_{(x_{i+1/2}, y_j)} \cdot h_y$, where

$$\begin{aligned} f_1 &= a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u = x \left[r_x(u) + d(y) \frac{\partial u}{\partial y} \right], \quad d(y) = \frac{1}{2} \rho \sigma(y) \sqrt{y}, \\ r_x(u) &\equiv a x \frac{\partial u}{\partial x} + b u, \quad a = 0.5y, \quad b = -y - 0.5 \rho \sigma'(y) \sqrt{y} - 0.25 \rho \frac{\sigma(y)}{\sqrt{y}}. \end{aligned}$$

Following the discussions in [1, 5, 9], *Case 1*, we approximate the “flux” $r_x(u)$ associated with respect to x by solving the following two-point BVP:

$$\begin{aligned} (a_{i+1/2,j} x v' + b_{i+1/2,j} v)' &= 0, \quad x \in (x_i, x_{i+1}), \\ v(x_i, y_j) &= u_{i,j}, \quad v(x_{i+1}, y_j) = u_{i+1,j}. \end{aligned}$$

By direct integration and approximation of $\partial u / \partial y$, for I_1 we obtain

$$I_1 \approx x_{i+1/2} \left(b_{i+1/2,j} \frac{x_{i+1}^{\alpha_{i,j}} u_{i+1,j} - x_i^{\alpha_{i,j}} u_{i,j}}{x_{i+1}^{\alpha_{i,j}} - x_i^{\alpha_{i,j}}} + d_{i+1/2,j} \frac{u_{i,j+1} + u_{i+1,j+1} - u_{i,j-1} - u_{i+1,j-1}}{4h_y} \right) h_y, \quad \alpha_{ij} = \frac{b_{i+1/2,j}}{a_{i+1/2,j}}.$$

In a similar way we find for the second integral for $i \geq 2$

$$I_2 \approx x_{i-1/2} \left(b_{i-1/2,j} \frac{x_i^{\alpha_{i-1,j}} u_{i,j} - x_{i-1}^{\alpha_{i-1,j}} u_{i-1,j}}{x_i^{\alpha_{i-1,j}} - x_{i-1}^{\alpha_{i-1,j}}} + d_{i-1/2,j} \frac{u_{i-1,j+1} + u_{i,j+1} - u_{i-1,j-1} - u_{i,j-1}}{4h_y} \right) h_y.$$

Note that the analysis in *Case 1* does not apply to approximation of the flux on $(0, x_1)$ because now the differential equation is *degenerate*. The approximation of I_2 for $i = 1$ (*Case 2*, [9]) requires the solution of the problem

$$\left(a_{1/2,j} x \frac{\partial v}{\partial x} + b_{1/2,j} v \right)' = C_2, \quad v(0, y_j) = g(0), \quad v(x_1, y_j) = u_{1,j},$$

which leads to

$$I_2 \approx x_{1/2} \left(\frac{1}{2} ((a_{1/2,j} + b_{1/2,j}) u_{1,j} - (a_{1/2,j} - b_{1/2,j}) u_{0,j}) + d_{1/2,j} \frac{u_{0,j+1} + u_{1,j+1} - u_{0,j-1} - u_{1,j-1}}{4h_y} \right) h_y.$$

For the *third* integral we obtain

$$I_3 \approx \left(\bar{b}_{i,j+1/2} \frac{y_{j+1}^{\bar{\alpha}_{i,j}} u_{i,j+1} - y_j^{\bar{\alpha}_{i,j}} u_{i,j}}{y_{j+1}^{\bar{\alpha}_{i,j}} - y_j^{\bar{\alpha}_{i,j}}} + \bar{d}_{i,j+1/2} \frac{u_{i+1,j} + u_{i+1,j+1} - u_{i-1,j} - u_{i-1,j+1}}{4h_x} \right) h_x, \quad \bar{\alpha}_{ij} = \frac{\bar{b}_{i+1/2,j}}{\bar{a}_{i+1/2,j}},$$

where $\bar{a} = 0.5\sigma^2(y)/2y$, $\bar{b} = \beta(y) - 0.5\rho\sigma(y)\sqrt{y} - \sigma(y)\sigma'(y)$, $\bar{d} = 0.5\rho x\sigma(y)\sqrt{y}$.

Next, for $2 \leq j \leq N_y - 1$, we have

$$I_4 \approx \left(\bar{b}_{i,j-1/2} \frac{y_j^{\bar{\alpha}_{i,j-1}} u_{i,j} - y_{j-1}^{\bar{\alpha}_{i,j-1}} u_{i,j-1}}{y_j^{\bar{\alpha}_{i,j-1}} - y_{j-1}^{\bar{\alpha}_{i,j-1}}} + \bar{d}_{i,j-1/2} \frac{u_{i+1,j} + u_{i+1,j-1} - u_{i-1,j} - u_{i-1,j-1}}{4h_x} \right) h_x.$$

For I_4 at $j = 1$ we get

$$I_4 \approx \left(\bar{d}_{i,1/2} \frac{u_{i+1,1} + u_{i+1,0} - u_{i-1,1} - u_{i-1,0}}{4h_x} + 0.5 \left((\bar{a}_{i,1/2} + \bar{b}_{i,1/2}) u_{i,1} - (\bar{a}_{i,1/2} - \bar{b}_{i,1/2}) u_{i,0} \right) \right) h_x.$$

In order to obtain semi-discrete equations at the mesh points $(x_i, 0)$ we integrate (1) on $\mathfrak{R}_{i,0} = [x_{i-1/2}, x_{i+1/2}] \times [0, y_{1/2}]$, $i = 1, \dots, N_x - 1$, Fig. 1:

$$\begin{aligned} \frac{\partial u_{i,0}}{\partial t} R_{i,0} &= \int_{\mathfrak{R}_{i,0}} \nabla \cdot (\mathbf{k}(u)) dx dy - p_{i,0} u_{i,0} R_{i,0}, \quad R_{i,0} = \frac{1}{2} h_x h_y, \\ \int_{\mathfrak{R}_{i,0}} \nabla \cdot (\mathbf{k}(u)) dx dy &= \int_{\partial \mathfrak{R}_{i,0}} \mathbf{k} \cdot \mathbf{n} ds = \int_{(x_{i+1/2}, 0)}^{(x_{i+1/2}, y_{1/2})} \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u \right) dy \\ &\quad - \int_{(x_{i-1/2}, 0)}^{(x_{i-1/2}, y_{1/2})} \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u \right) dy \\ &\quad + \int_{(x_{i-1/2}, y_{1/2})}^{(x_{i+1/2}, y_{1/2})} \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} + b_2 u \right) dx - \int_{(x_{i-1/2}, 0)}^{(x_{i+1/2}, 0)} \left(a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} + b_2 u \right) dx \\ &= I_1^d - I_2^d + I_3^d - I_4^d. \end{aligned}$$

For I_1^d we get

$$I_1^d \approx \left(a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} + b_1 u \right) \Big|_{(x_{i+1/2}, 0)} \frac{h_y}{2} \approx \frac{h_y}{2} b_1(x_{i+1/2}, 0) \frac{u_{i+1,0} + u_{i,0}}{2}.$$

Similarly,

$$\begin{aligned} I_2^d &\approx \frac{h_y}{2} b_1(x_{i-1/2}, 0) \frac{u_{i,0} + u_{i-1,0}}{2}, \\ I_3^d &\approx \left(\bar{d}_{i,1/2} \frac{u_{i+1,1} + u_{i+1,0} - u_{i-1,1} - u_{i-1,0}}{4h_x} + \frac{1}{2} \left((\bar{a}_{i,1/2} + \bar{b}_{i,1/2}) u_{i,1} - (\bar{a}_{i,1/2} - \bar{b}_{i,1/2}) u_{i,0} \right) \right) h_x, \\ I_4^d &\approx h_x b_2(x_i, 0) u_{i,0}. \end{aligned}$$

Finally, on the base of all constructions above, we obtain the following system of ODEs:

$$\begin{aligned} \frac{\partial u_{i,0}}{\partial t} R_{i,0} - e_{i,0,i-1,0} u_{i-1,0} - e_{i,0,i-1,1} u_{i-1,1} + e_{i,0,i,0} u_{i,0} - e_{i,0,i,1} u_{i,1} \\ - e_{i,0,i+1,0} u_{i+1,0} - e_{i,0,i+1,1} u_{i+1,1} = 0, \quad \text{for } i = 1, 2, \dots, N_x - 1; \end{aligned}$$

$$\begin{aligned}
& \frac{\partial u_{ij}}{\partial t} R_{ij} - e_{i,j,i-1,j-1} u_{i-1,j-1} - e_{i,j,i-1,j} u_{i-1,j} - e_{i,j,i-1,j+1} u_{i-1,j+1} \\
& - e_{i,j,i,j-1} u_{i,j-1} + e_{i,j,i,j} u_{i,j} - e_{i,j,i,j+1} u_{i,j+1} - e_{i,j,i+1,j-1} u_{i+1,j-1} \\
& - e_{i,j,i+1,j} u_{i+1,j} - e_{i,j,i+1,j+1} u_{i+1,j+1} = 0 \quad (6)
\end{aligned}$$

for $i = 1, 2, \dots, N_x - 1$, $j = 1, 2, \dots, N_y - 1$ and $u_{0,j} = g(0)$, $u_{N_x,j} = g(X)$, $u_{i,N_y} = g(x_i)$, $i = 1, 2, \dots, N_x - 1$, $j = 0, 1, \dots, N_y$. The coefficients are defined by

$$\begin{aligned}
& e_{1,j,0,j\pm 1} = \mp 0.25 (x_{1/2} d_{1/2,j} + \bar{d}_{1,j\pm 1/2}), \quad e_{1,j,0,j} = 0.5 h_y x_{1/2} (a_{1/2,j} - b_{1/2,j}) \\
& + \frac{\bar{d}_{1,j-1/2} - \bar{d}_{1,j+1/2}}{4}, \quad e_{1,j,1,j-1} = \frac{h_x \bar{b}_{1,j-1/2} y_{j-1}^{\bar{\alpha}_{1,j-1}}}{y_j^{\bar{\alpha}_{1,j-1}} - y_{j-1}^{\bar{\alpha}_{1,j-1}}} + \frac{x_{1/2} d_{1/2,j} - x_{3/2} d_{3/2,j}}{4}, \\
& e_{1,j,1,j+1} = \frac{h_x \bar{b}_{1,j+1/2} y_{j+1}^{\bar{\alpha}_{1,j}}}{y_{j+1}^{\bar{\alpha}_{1,j}} - y_j^{\bar{\alpha}_{1,j}}} + \frac{x_{3/2} d_{3/2,j} - x_{1/2} d_{1/2,j}}{4}, \quad e_{1,j,1,j} = \frac{h_y x_{3/2} b_{3/2,j} x_1^{\alpha_{1,j}}}{x_2^{\alpha_{1,j}} - x_1^{\alpha_{1,j}}} \\
& + \frac{h_y x_{1/2} (a_{1/2,j} + b_{1/2,j})}{2} + \frac{h_x \bar{b}_{1,j+1/2} y_j^{\bar{\alpha}_{1,j}}}{y_{j+1}^{\bar{\alpha}_{1,j}} - y_j^{\bar{\alpha}_{1,j}}} + \frac{h_x \bar{b}_{1,j-1/2} y_j^{\bar{\alpha}_{1,j-1}}}{y_j^{\bar{\alpha}_{1,j-1}} - y_{j-1}^{\bar{\alpha}_{1,j-1}}} + h_x h_y p_{1,j}, \\
& e_{1,j,2,j} = \frac{h_y x_{3/2} b_{3/2,j} x_2^{\alpha_{1,j}}}{x_2^{\alpha_{1,j}} - x_1^{\alpha_{1,j}}} + \frac{\bar{d}_{1,j+1/2} - \bar{d}_{1,j-1/2}}{4}, \\
& e_{1,j,2,j\pm 1} = \pm 0.25 (x_{3/2} d_{3/2,j} + \bar{d}_{1,j\pm 1/2}), \quad j = 2, \dots, N_y - 1; \\
& e_{i,1,i\pm 1,0} = \mp 0.25 (x_{i\pm 1/2} d_{i\pm 1/2,1} + \bar{d}_{i,1/2}), \\
& e_{i,1,i,0} = 0.25 (x_{i-1/2} d_{i-1/2,1} - x_{i+1/2} d_{i+1/2,1} + 2h_x (\bar{a}_{i,1/2} - \bar{b}_{i,1/2})), \\
& e_{i,1,i-1,1} = \frac{h_y x_{i-1/2} b_{i-1/2,1} x_{i-1}^{\alpha_{i-1,1}}}{x_i^{\alpha_{i-1,1}} - x_{i-1}^{\alpha_{i-1,1}}} + \frac{\bar{d}_{i,1/2} - \bar{d}_{i,3/2}}{4}, \\
& e_{i,1,i+1,1} = \frac{h_y x_{i+1/2} b_{i+1/2,1} x_{i+1}^{\alpha_{i,1}}}{x_{i+1}^{\alpha_{i,1}} - x_i^{\alpha_{i,1}}} + \frac{\bar{d}_{i,3/2} - \bar{d}_{i,1/2}}{4}, \\
& e_{i,1,i,1} = \frac{h_y x_{i+1/2} b_{i+1/2,1} x_i^{\alpha_{i,1}}}{x_{i+1}^{\alpha_{i,1}} - x_i^{\alpha_{i,1}}} + \frac{h_y x_{i-1/2} b_{i-1/2,1} x_i^{\alpha_{i-1,1}}}{x_i^{\alpha_{i-1,1}} - x_{i-1}^{\alpha_{i-1,1}}} + \frac{h_x \bar{b}_{i,3/2} y_1^{\bar{\alpha}_{i,1}}}{y_2^{\bar{\alpha}_{i,1}} - y_1^{\bar{\alpha}_{i,1}}} \\
& + \frac{h_x (\bar{a}_{i,1/2} + \bar{b}_{i,1/2})}{2} + h_x h_y p_{i,1}, \quad e_{i,1,i,2} = \frac{h_x \bar{b}_{i,3/2} y_2^{\bar{\alpha}_{i,1}}}{y_2^{\bar{\alpha}_{i,1}} - y_1^{\bar{\alpha}_{i,1}}} + \frac{x_{i+1/2} \bar{d}_{i+1/2,1}}{4} \\
& - \frac{x_{i-1/2} \bar{d}_{i-1/2,1}}{4}, \quad e_{i,1,i\pm 1,2} = \pm 0.25 (x_{i\pm 1/2} d_{i\pm 1/2,1} + \bar{d}_{i,3/2}), \quad i = 2, \dots, N_x - 1; \\
& e_{i,0,i,0} = 0.25 h_y (b_1(x_{i-1/2}, 0) - b_1(x_{i+1/2}, 0)) + 0.5 h_x (\bar{a}_{i,1/2} - \bar{b}_{i,1/2} + h_y p_{i,0}) \\
& + h_x b_2(x_i, 0), \quad e_{i,0,i\pm 1,0} = \pm 0.25 (h_y b_1(x_{i\pm 1/2}, 0) + \bar{d}_{i,1/2}), \\
& e_{i,0,i,1} = 0.5 h_x (\bar{a}_{i,1/2} + \bar{b}_{i,1/2}), \quad e_{i,0,i\pm 1,1} = \pm 0.25 \bar{d}_{i,1/2}, \quad i = 1, 2, \dots, N_x - 1; \\
& e_{1,1,0,0} = 0.25 (\bar{d}_{1,1/2} + x_{1/2} d_{1/2,1}), \quad e_{1,1,0,2} = -0.25 (\bar{d}_{1,3/2} + x_{1/2} d_{1/2,1}), \\
& e_{1,1,0,1} = 0.25 (\bar{d}_{1,1/2} - \bar{d}_{1,3/2} + 2h_y x_{1/2} (a_{1/2,1} - b_{1/2,1})), \\
& e_{1,1,1,0} = 0.25 (x_{1/2} d_{1/2,1} - x_{3/2} d_{3/2,1} + 2h_x (\bar{a}_{1,1/2} - \bar{b}_{1,1/2})),
\end{aligned}$$

$$\begin{aligned}
e_{1,1,1,1} &= \frac{h_y x_{3/2} b_{3/2,1} x_1^{\alpha_{1,1}}}{x_2^{\alpha_{1,1}} - x_1^{\alpha_{1,1}}} + \frac{h_x \bar{b}_{1,3/2} y_1^{\bar{\alpha}_{1,1}}}{y_2^{\bar{\alpha}_{1,1}} - y_1^{\bar{\alpha}_{1,1}}} + h_x h_y p_{1,1} \\
&\quad + 0.5 \left[h_y x_{1/2} (a_{1/2,1} + b_{1/2,1}) + h_x (\bar{a}_{1,1/2} + \bar{b}_{1,1/2}) \right], \\
e_{1,1,1,2} &= \frac{h_x \bar{b}_{1,3/2} y_2^{\bar{\alpha}_{1,1}}}{y_2^{\bar{\alpha}_{1,1}} - y_1^{\bar{\alpha}_{1,1}}} + \frac{x_{3/2} d_{3/2,1} - x_{1/2} d_{1/2,1}}{4}, \\
e_{1,1,2,0} &= -0.25 (x_{3/2} d_{3/2,1} + \bar{d}_{1,1/2}), e_{1,1,2,2} = 0.25 (x_{3/2} d_{3/2,1} + \bar{d}_{1,3/2}), \\
e_{1,1,2,1} &= \frac{h_y x_{3/2} b_{3/2,1} x_2^{\alpha_{1,1}}}{x_2^{\alpha_{1,1}} - x_1^{\alpha_{1,1}}} + \frac{\bar{d}_{1,3/2} - \bar{d}_{1,1/2}}{4}; \\
e_{i,j,i-1,j\pm 1} &= \mp 0.25 (x_{i-1/2} d_{i-1/2,j} + \bar{d}_{i,j\pm 1/2}), \\
e_{i,j,i-1,j} &= \frac{h_y x_{i-1/2} b_{i-1/2,j} x_i^{\alpha_{i-1,j}}}{x_i^{\alpha_{i-1,j}} - x_{i-1}^{\alpha_{i-1,j}}} + \frac{\bar{d}_{i,j-1/2} - \bar{d}_{i,j+1/2}}{4}, \\
e_{i,j,i,j-1} &= \frac{h_x \bar{b}_{i,j-1/2} y_j^{\bar{\alpha}_{i,j-1}}}{y_j^{\bar{\alpha}_{i,j-1}} - y_{j-1}^{\bar{\alpha}_{i,j-1}}} + \frac{x_{i-1/2} d_{i-1/2,j} - x_{i+1/2} d_{i+1/2,j}}{4}, \\
e_{i,j,i,j} &= \frac{h_y x_{i+1/2} b_{i+1/2,j} x_i^{\alpha_{i,j}}}{x_{i+1}^{\alpha_{i,j}} - x_i^{\alpha_{i,j}}} + \frac{h_y x_{i-1/2} b_{i-1/2,j} x_i^{\alpha_{i-1,j}}}{x_i^{\alpha_{i-1,j}} - x_{i-1}^{\alpha_{i-1,j}}} \\
&\quad + \frac{h_x \bar{b}_{i,j+1/2} y_j^{\bar{\alpha}_{i,j}}}{y_{j+1}^{\bar{\alpha}_{i,j}} - y_j^{\bar{\alpha}_{i,j}}} + \frac{h_x \bar{b}_{i,j-1/2} y_j^{\bar{\alpha}_{i,j-1}}}{y_j^{\bar{\alpha}_{i,j-1}} - y_{j-1}^{\bar{\alpha}_{i,j-1}}} + h_x h_y p_{i,j}, \\
e_{i,j,i,j+1} &= \frac{h_x \bar{b}_{i,j+1/2} y_{j+1}^{\bar{\alpha}_{i,j}}}{y_{j+1}^{\bar{\alpha}_{i,j}} - y_j^{\bar{\alpha}_{i,j}}} + \frac{x_{i+1/2} d_{i+1/2,j} - x_{i-1/2} d_{i-1/2,j}}{4}, \\
e_{i,j,i+1,j} &= \frac{h_y x_{i+1/2} b_{i+1/2,j} x_{i+1}^{\alpha_{i,j}}}{x_{i+1}^{\alpha_{i,j}} - x_i^{\alpha_{i,j}}} + \frac{\bar{d}_{i,j+1/2} - \bar{d}_{i,j-1/2}}{4}, e_{i,j,i+1,j\pm 1} \\
&= \pm 0.25 (x_{i+1/2} d_{i+1/2,j} + \bar{d}_{i,j\pm 1/2}), i = 2, \dots, N_x - 1, j = 2, \dots, N_y - 1.
\end{aligned}$$

In conclusion we may write the following result:

Theorem 1. The semi-discretization (6) is consistent with equation (1) and the truncation error is of order $O(h_x + h_y)$.

4 Full Discretization

The ODEs above form an $(N_x - 1)N_y \times (N_x - 1)N_y$ linear system for

$$\mathbf{u} = (u_{1,0}, \dots, u_{1,N_y-1}, u_{2,0}, \dots, u_{2,N_y-1}, \dots, u_{N_x-1,0}, \dots, u_{N_x-1,N_y-1})^T$$

with $u_{0,j}(t), u_{i,N_y}(t), u_{N_x,j}(t), i = 1, \dots, N_x - 1, j = 0, \dots, N_y$ being equal to the right hand side of the given Dirichlet boundary conditions. Let

$$\begin{aligned}
\mathbf{E}_{i,0} &= (0, \dots, 0, -e_{i,0,i-1,0}, -e_{i,0,i-1,1}, 0, \dots, 0, e_{i,0,i,0}, -e_{i,0,i,1}, 0, \dots, 0, \\
&\quad -e_{i,0,i+1,0}, -e_{i,0,i+1,1}, 0, \dots, 0),
\end{aligned}$$

$\mathbf{E}_{i,j} = (0, \dots, 0, -e_{i,j,i-1,j-1}, -e_{i,j,i-1,j}, -e_{i,j,i-1,j+1}, 0, \dots, 0, -e_{i,j,i,j-1}, e_{i,j,i,j}, -e_{i,j,i,j+1}, 0, \dots, 0, -e_{i,j,i+1,j-1}, -e_{i,j,i+1,j}, -e_{i,j,i+1,j+1}, 0, \dots, 0)$
for $i = 1, 2, \dots, N_x - 1, j = 1, 2, \dots, N_y - 1$. Now, the ODEs takes the form

$$\frac{\partial u_{i,j}(t)}{\partial t} R_{i,j} + \mathbf{E}_{i,j}(t) \mathbf{u}(t) = 0 \quad (7)$$

for $i = 1, 2, \dots, N_x - 1, j = 0, 1, \dots, N_y - 1$. To discretize this system we let t_i ($i = 0, 1, \dots, K$) be a set of partition points in $[0, T]$, satisfying $0 = t_0 < t_1 < \dots < t_K = T$. Then, we apply the two-level implicit-stepping method with a splitting parameter $\theta \in [0, 1]$ to (6) to yield

$$\frac{u_{i,j}^{k+1} - u_{i,j}^k}{\tau_k} R_{i,j} + \theta \mathbf{E}_{i,j}^{k+1} \mathbf{u}^{k+1} + (1 - \theta) \mathbf{E}_{i,j}^k \mathbf{u}^k = 0$$

for $k = 0, 1, 2, \dots, K - 1$, where $\tau_k = t_{k+1} - t_k > 0$, $\mathbf{E}_{i,j}^k = \mathbf{E}_{i,j}(t_k)$ and $\mathbf{u}^k$ denotes the approximation of u at $t = t_k$. Let $\mathbf{E}^k$ be the $(N_x - 1)N_y \times (N_x - 1)N_y$ matrix given by

$$\mathbf{E}^k = (\mathbf{E}_{1,0}^k, \mathbf{E}_{1,1}^k, \dots, \mathbf{E}_{N_x-1,0}^k, \mathbf{E}_{N_x-1,1}^k, \dots, \mathbf{E}_{N_x-1,N_y-1}^k)^T.$$

Then the above system can be rewritten as

$$(\theta \mathbf{E}^{k+1} + \mathbf{G}^k) \mathbf{u}^{k+1} = [\mathbf{G}^k - (1 - \theta) \mathbf{E}^k] \mathbf{u}^k \quad (8)$$

for $k = 0, 1, \dots, K - 1$, where $\mathbf{G}^k = \text{diag}(R_{1,0}/\tau_k, \dots, R_{N_x-1,N_y-1}/\tau_k)$ is an $(N_x - 1)N_y \times (N_x - 1)N_y$ diagonal matrix. When $\theta = 0.5$ the time-stepping scheme becomes the Crank-Nicolson scheme and when $\theta = 1$ it is the implicit scheme. Both of these schemes are unconditionally stable and they are of second- and first-order accuracy respectively with respect to time [9].

5 Numerical Experiments

Numerical experiments were performed in order to examine the properties of the constructed scheme. We approximately solve model problem with known analytical solution $u_{ex}(x, y, t) = x \exp(-ty)$. We choose this function because its character is similar to the character of the exact solution of the problem under consideration. The coefficients are $\sigma(y) = c\sqrt{y}$, $\beta(y) = a(b - y)$ (CIR-model). The other data are the following: $X = Y = T = 1$, $\rho = 0.5$, $a = 0.55$, $b = 0.035$, $c = 0.39$ [2]. Results from computational experiments concerning the error and the rate of convergence (RC) with respect to space are presented in Table 1. Everywhere the calculations are performed with constant time step $\tau = 2^{-12} = 0.000244140625$. We chose this small time step because it has no influence on the error of the numerical results. The rate of convergence (RC) is calculated using double mesh principle

$$RC = \log_2(ER^N / ER^{2N}), \quad ER^N = \|u_{ex}^N - u^N\|,$$

where $\|\cdot\|$ is the mesh C -norm or L_2 -norm, u_{ex}^N and u^N are respectively the exact solution and the numerical solution computed at the mesh, $N = N_x = N_y$.

Table 1. Crank-Nicolson scheme results (Dirichlet boundary conditions at $X=1$ and $Y=1$)

$N_x \times N_y$	C -norm of error	C -norm RC	L_2 -norm error	L_2 -norm RC
4×4	2.835 E-2	-	5.217 E-3	-
8×8	1.694 E-2	0.74	2.108 E-3	1.30
16×16	9.126 E-3	0.90	7.596 E-4	1.48
32×32	4.637 E-3	0.98	2.678 E-4	1.51
64×64	2.184 E-3	1.08	9.573 E-5	1.49
128×128	1.026 E-3	1.09	3.431 E-5	1.48

6 Conclusions

In this paper, we derived a finite volume difference scheme with a fitting technique for the numerical solution of a 2D Black-Scholes equation, the CIR model as a typical example. The derivation of the scheme and the numerical experiments show that it is first order accurate in space. In future work we plan to study more in detail the monotone properties and the convergence of the difference scheme in strong and Sobolev discrete norms. Along the boundaries $x = X$ and $y = Y$ we used Dirichlet boundary conditions. But this question requires a special investigation. We plan to apply the energy method of Godunov to derive boundary conditions at outer boundaries for which the problem will be well-posed on a finite domain.

Acknowledgment. The first author is supported by the Sofia University Foundation under Grant No 196/2010 and the second author is supported by the Project Bg-Sk-203.

References

1. Chernogorova, T., Valkov, R.: A computational scheme for a problem in the zero-coupon bond pricing. Amer. Inst. of Phys. (in press)
2. Ekstrom, E., Lotstedt, P., Tysk, J.: Boundary values and finite difference methods for the single-factor term structure equation. Appl. Math. Finance 16, 252–259 (2009)
3. Ekstrom, E., Tysk, J.: The Black-Scholes equation in stochastic volatility models. J. Math. Anal. Appl. 368, 498–507 (2010)
4. Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Finan. Stud. 6, 327–343 (1993)
5. Huang, C.-S., Hung, C.-H., Wang, S.: A fitted finite volume method for the valuation of options on assets with stochastic volatilities. Computing 77, 297–320 (2006)
6. Lions, P.-L., Musiela, M.: Correlations and bounds for stochastic volatility models. Ann. Inst. H. Poincaré Anal. Non Linéaire 24, 1–16 (2007)
7. Oleinik, O.A., Radkevich, E.V.: Second Order Equations with Nonnegative Characteristic Form. Plenum Press, New York (1973)
8. Thomas, J.W.: Numerical Partial Differential Equations. Springer, Berlin (1995)
9. Wang, S.: A novel fitted finite volume method for Black-Scholes equation governing option pricing. IMA J. of Numer. Anal. 24, 699–720 (2004)

On the Numerical Simulation of Unsteady Solutions for the 2D Boussinesq Paradigm Equation

Christo I. Christov¹, Natalia Kolkovska², and Daniela Vasileva²

¹ Dept. of Mathematics, P.O. Box 41010, Lafayette, LA, 70504-1010, USA
`christov@louisiana.edu`

² Institute of Mathematics and Informatics, Bulgarian Acad. Sci.,
Acad. G. Bonchev str., bl.8, 1113 Sofia, Bulgaria
`{natali,vasileva}@math.bas.bg`

Abstract. For the solution of the 2D Boussinesq Paradigm Equation (BPE) an implicit, unconditionally stable difference scheme with second order truncation error in space and time is designed. Two different asymptotic boundary conditions are implemented: the trivial one, and a condition that matches the expected asymptotic behavior of the profile at infinity. The available in the literature solutions of BPE of type of stationary localized waves are used as initial conditions for different phase speeds and their evolution is investigated numerically. We find that, the solitary waves retain their identity for moderate times; for larger times they either transform into diverging propagating waves or blow-up.

1 Introduction

Boussinesq equation (BE) [1] is the first model for surface waves in shallow fluid layer that accounts for both nonlinearity and dispersion. The balance between the steepening effect of the nonlinearity and the flattening effect of the dispersion maintains the shape of the wave. In the 60s it was discovered that these permanent waves can behave in many instances as particles in 1D and they were called *solitons* by Zabusky and Kruskal [2]. It is of crucial importance to investigate also the 2D case, because of the different phenomenology and the practical importance. The accurate derivation of the Boussinesq system combined with an approximation, that reduces the full model to a single equation, leads to the Boussinesq Paradigm Equation (BPE) [3]:

$$u_{tt} = \Delta [u - F(u) + \beta_1 u_{tt} - \beta_2 \Delta u], \quad F(u) := \alpha u^2, \quad (1)$$

where u is the surface elevation, $\beta_1, \beta_2 > 0$ are two dispersion coefficients, and $\alpha > 0$ is an amplitude parameter. The main difference of (1) from BE is the presence of a term proportional to $\beta_1 \neq 0$ called “rotational inertia”. Note that here we have changed the sign of the nonlinear term for the sake of the presentation.

It has been recently shown that the 2D BPE admits stationary translating localized solutions as well [4–7]. Even though no exact analytical formulas are available, those solutions can be accessed using either finite differences, perturbation technique, or Galerkin spectral method. However, virtually nothing is known about the dynamic properties of these solutions and their structural stability, i.e., what is their behavior when used as initial conditions for time-dependent computations of the BPE. The first results on this problem are reported in the pioneering work [8], but in order to investigate further the time evolution of the localized solutions, alternative techniques for Eq. (1) have to be developed.

2 Numerical Method for Solving BPE

In order to devise a numerical time-stepping procedure for Eq. (1), we set

$$v(x, y, t) := u - \beta_1 \Delta u. \quad (2a)$$

Upon substituting it in Eq. (1) we get the following equation for v

$$v_{tt} = \frac{\beta_2}{\beta_1} \Delta v + \frac{\beta_1 - \beta_2}{\beta_1^2} (u - v) - \Delta F(u). \quad (2b)$$

Now the system consists of an elliptic equation for u , Eq. (2a), and a hyperbolic equation for v : Eq. (2b). The system is inextricable coupled, because the function u is involved in the equation for v , and vice versa.

The following implicit time stepping can be designed for the system (2)

$$\begin{aligned} \frac{v_{ij}^{n+1} - 2v_{ij}^n + v_{ij}^{n-1}}{\tau^2} &= \frac{\beta_2}{2\beta_1} \Lambda [v_{ij}^{n+1} + v_{ij}^{n-1}] \\ &+ \frac{\beta_1 - \beta_2}{2\beta_1^2} [u_{ij}^{n+1} - v_{ij}^{n+1} + u_{ij}^{n-1} - v_{ij}^{n-1}] - \Lambda F(u_{ij}^n), \end{aligned} \quad (3a)$$

$$u_{ij}^{n+1} - \beta_1 \Lambda u_{ij}^{n+1} = v_{ij}^{n+1}, \quad i = 0, \dots, N_x + 1, \quad j = 0, \dots, N_y + 1. \quad (3b)$$

Here τ is the time increment, and $\Lambda = \Lambda^{xx} + \Lambda^{yy}$ stands for the difference approximation of the Laplace operator Δ on a non-uniform grid, for example

$$\Lambda^{xx} \phi_{ij} = \frac{2\phi_{i-1j}}{h_{i-1}^x(h_i^x + h_{i-1}^x)} - \frac{2\phi_{ij}}{h_i^x h_{i-1}^x} + \frac{2\phi_{i+1j}}{h_i^x(h_i^x + h_{i+1}^x)} = \frac{\partial^2 \phi}{\partial x^2} \Big|_{ij} + O(|h_i^x - h_{i-1}^x|).$$

For a smooth distribution of the nonuniform grid (as the one considered here) one has

$$O(|h_i^x - h_{i-1}^x|) \approx \frac{\partial h^x}{\partial x} O(|h_{i-1}|^2) = O(|h_{i-1}|^2).$$

Respectively, the values of the sought functions at the $(n-1)$ -st and n -th time stages are considered as known when computing the $(n+1)$ -st stage. Thus, we have two *coupled* equations for the two unknown grid functions u_{ij}^{n+1} , v_{ij}^{n+1} and use the following non-uniform grid in the x -direction

$$x_i = \sinh[\hat{h}_x(i - n_x)], \quad x_{N_x+1-i} = -x_i, \quad i = n_x + 1, \dots, N_x + 1, \quad x_{n_x} = 0,$$

where N_x is an odd number, $n_x = (N_x + 1)/2$, $\hat{h}_x = D_x/N_x$, and D_x is selected in a manner to have large enough computational region. The grid in the y -direction is defined in the same way.

The unconditional stability of the scheme can be shown in a way, very similar to [9], where numerical experiments in the 1D case with the analogue of the scheme (3), confirm the findings in the literature (see, e.g. [10]) that the BPE solitons preserve their shape for all times and even after interaction.

In the simplest approximation, the boundary conditions can be set equal to zero, because of the localization of the wave profile. This forms the first set of b.c.'s used in the present work. However, the decay at infinity of the stationary propagating 2D Boussinesq solitons is second-order algebraic (see [4, 6]), which requires really large computational box in order that the solution in the main part of the region (far from the boundaries) is not adversely influenced. Thus, the second set of b.c.'s used in the present work are the asymptotic boundary conditions formulated in [7]

$$x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} \approx -2u, \quad x \frac{\partial v}{\partial x} + y \frac{\partial v}{\partial y} \approx -2v, \quad \sqrt{x^2 + y^2} \gg 1. \quad (4)$$

We chose the following approximation for Eq. (4)₁ at the numerical infinities:

$$u_{i,N_y+1}^{n+1} = u_{i,N_y-1}^{n+1} + \frac{h_{N_y}^y + h_{N_y-1}^y}{y_{N_y}} \left[-2u_{i,N_y}^{n+1} - \frac{x_i}{h_i^x + h_{i-1}^x} (u_{i+1,N_y}^{n+1} - u_{i-1,N_y}^{n+1}) \right],$$

$$u_{N_x+1,j}^{n+1} = u_{N_x-1,j}^{n+1} + \frac{h_{N_x}^x + h_{N_x-1}^x}{x_{N_x}} \left[-2u_{N_x,j}^{n+1} - \frac{y_j}{h_j^y + h_{j-1}^y} (u_{N_x,j+1}^{n+1} - u_{N_x,j-1}^{n+1}) \right],$$

$i = 0, \dots, N_x$, $j = 0, \dots, N_y$. The implementation of Eq. (4)₂ is the same.

The initial conditions are created using the best-fit approximation provided in [6], and already used in [8]. The coupled system of equations (3) is solved by the Bi-Conjugate Gradient Stabilized Method with ILU preconditioner [11].

3 Numerical Experiments

Denote by $u^s(x, y; c)$ the best-fit approximation of the stationary translating (with speed c) localized solutions, obtained in [6]

$$u^s(x, y; c) = f(x, y) + c^2 [(1 - \beta_1)g_a(x, y) + \beta_1 g_b(x, y)] \\ + c^2 [(1 - \beta_1)h_1(x, y) + \beta_1 h_2(x, y)] \cos[2 \arctan(y/x)],$$

where the formulas for the functions f , g_a , g_b may be found in [6]. For $t = 0$, the first initial condition is obvious: $u(x, y, 0) = u^s(x, y; c)$, the second initial condition may be chosen as one of the following

$$\partial u / \partial t = -c \partial u^s / \partial y \quad \text{or} \quad u(x, y, -\tau) = u^s(x, y + c\tau; c), \quad (5)$$

and (5)₁ is approximated as $\frac{u_{ij}^1 - u_{ij}^{-1}}{2\tau} = -c \frac{\partial u^s}{\partial y}(x_i, y_j)$.

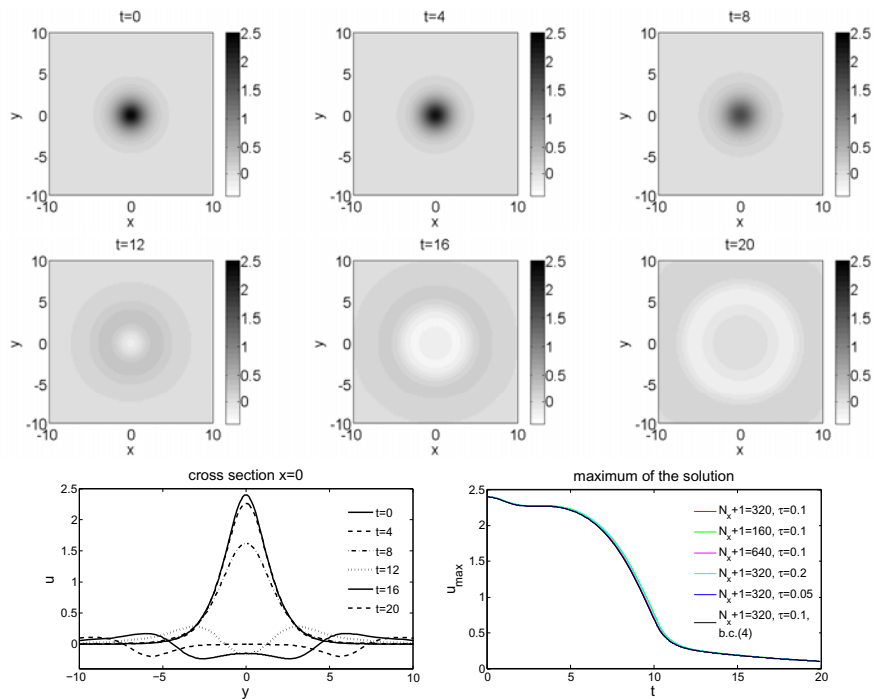


Fig. 1. Evolution of the solution for $c = 0$, the evolution of the cross-section at $x = 0$ and the values of the maximum

Table 1. The maximum of the solution, convergence in space and time, $c = 0$

		$t = 4$			$t = 8$			$t = 12$		
τ	$N_x + 1$	$u_{\max}$	$\Delta u_{\max}$	l	$u_{\max}$	$\Delta u_{\max}$	l	$u_{\max}$	$\Delta u_{\max}$	l
with second IC according to (5) ₁										
0.1	160	2.27122			1.64704			2.87575e-1		
0.1	320	2.26475	6.47e-3		1.60553	4.15e-2		2.80298e-1	7.28e-3	
0.1	640	2.26314	1.62e-3	2.0	1.59531	1.02e-2	2.0	2.78648e-1	1.65e-3	2.1
0.1	320	2.26475			1.60553			2.80298e-1		
0.05	320	2.26464	1.17e-4		1.60238	3.14e-3		2.79847e-1	4.51e-4	
0.025	320	2.26461	3.00e-5	2.0	1.60159	7.89e-4	2.0	2.79736e-1	1.11e-4	2.0
with second IC according to (5) ₂										
0.1	160	2.27016			1.62990			2.84523e-1		
0.1	320	2.26350	6.66e-3		1.58771	4.22e-2		2.77527e-1	6.996e-3	
0.1	640	2.26184	1.66e-3	2.0	1.57733	1.04e-2	2.0	2.75936e-1	1.591e-3	2.1
0.0125	320	2.26444			1.59915			2.79355e-1		
0.00625	320	2.26452	-7.80e-5		1.60022	-1.08e-3		2.79524e-1	-1.69e-4	
0.003125	320	2.26456	-4.00e-5	1.0	1.60077	-5.49e-4	1.0	2.79611e-1	-8.67e-5	1.0

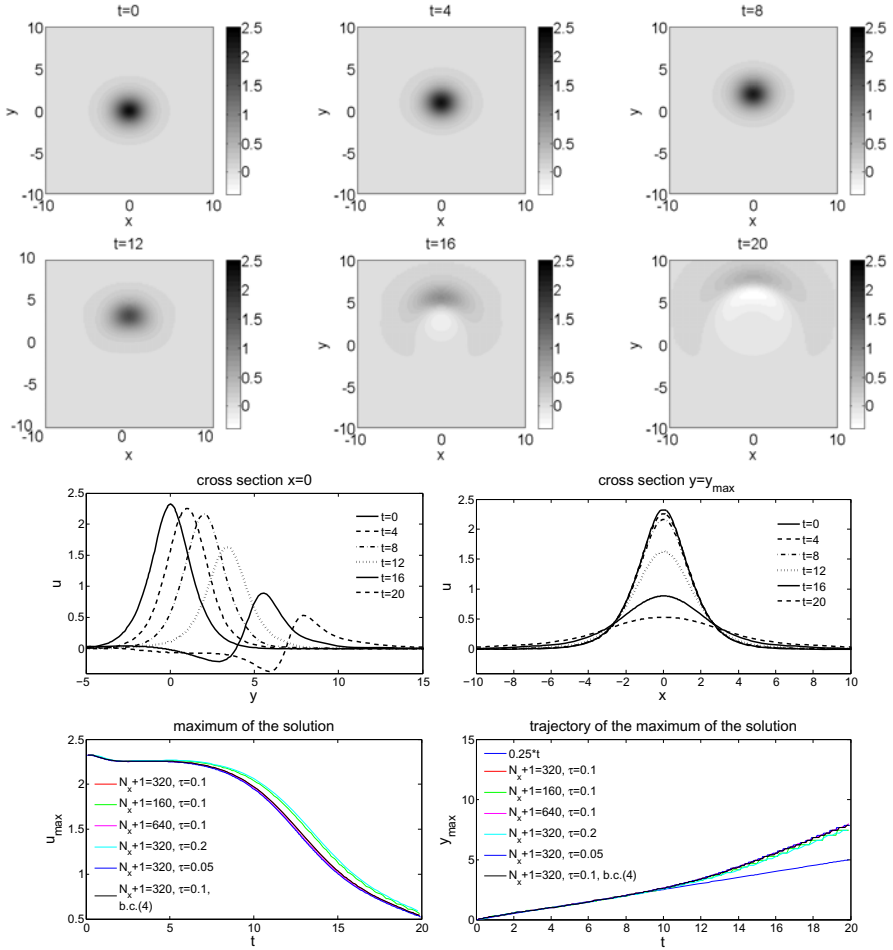


Fig. 2. Evolution of the solution for $c = 0.25$, evolution of the cross sections at $x = 0$ and $y = y_{\max}$, the maximum $u(0, y_{\max})$, and the trajectory of the maximum

The solutions for $\beta_1 = 3$, $\beta_2 = 1$, $\alpha = 1$ are computed on three different grids in the region $x, y \in [-50, 50]$ (with 161×161 , 321×321 and 641×641 grid points), with at least three different time increments ($\tau = 0.2, 0.1$ and 0.05), and using either the trivial boundary conditions or the conditions (4).

Example 1. First, we present the results for the case $c = 0$, when the profile of the initial condition is a standing soliton. As it is seen in Fig. 1, the nonlinearity is not strong enough and after $t \geq 4$ the solution cannot keep the form, and eventually transforms into a propagating cylindrical wave, similar to the one generated on a water surface when an object is dropped into it (note, the sign of the solution is reversed in BPE (1)). The ‘longitudinal’ cross-section of the

Table 2. The maximum of the solution, convergence in space and time, $c = 0.25$

		$t = 4$			$t = 8$			$t = 12$		
τ	$N_x + 1$	$u_{\max}$	$\Delta u_{\max}$	l	$u_{\max}$	$\Delta u_{\max}$	l	$u_{\max}$	$\Delta u_{\max}$	l
with second IC according to (5) ₁										
0.1	160	2.261156			2.191684			1.725273		
0.1	320	2.257642	3.51e-3		2.165738	2.59e-2		1.639348	8.59e-2	
0.1	640	2.256689	9.53e-4	1.9	2.158619	7.12e-3	1.9	1.619535	1.98e-2	2.1
0.2	320	2.268606			2.226354			1.848499		
0.1	320	2.257642	1.10e-2		2.165738	6.06e-2		1.639348	2.09e-1	
0.05	320	2.254871	2.77e-3	2.0	2.148196	1.75e-2	1.8	1.588800	5.05e-2	2.0
with second IC according to (5) ₂										
0.1	160	2.261550			2.189987			1.718885		
0.1	320	2.256804	4.75e-3		2.156155	3.38e-2		1.609205	1.10e-1	
0.1	640	2.255469	1.34e-3	1.8	2.147008	9.15e-3	1.9	1.584249	2.50e-2	2.1
0.2	320	2.264348			2.195763			1.734455		
0.1	320	2.256804	7.54e-3		2.156155	3.96e-2		1.609205	1.25e-1	
0.05	320	2.254958	1.85e-3	2.0	2.146491	9.66e-3	2.0	1.583778	2.54e-2	2.3

solution at $x = 0$ for a couple of moments of time and the values of the maximum of the solution as function of time are also shown in Fig. 1. The behaviour of the solution is the same on all grids, for all times steps, and does not depend on the type of the boundary conditions used (the trivial one or (4)).

For $t = 4, 8, 12$ the computed maximum of the solution $u_{\max}$, the difference $\Delta u_{\max} := u_{\max}^{\text{prev}} - u_{\max}$ (subscript ‘prev’ denotes the previous row in the table), and the rate of convergence $l = \log_2(|u_{\max}^{\text{prev}} - u_{\max}^{\text{prev,prev}}|/|u_{\max} - u_{\max}^{\text{prev}}|)$, are shown in Table 1. It is seen that when the second initial condition is taken according to (5)₁ the method is second order accurate in space and time. When the second initial condition is posed at $t = -\tau$ (i.e., (5)₂ is used), the method is only first order accurate in time, but this does not change significantly the behaviour of the solution, because the effect is localized near the initial moment of time.

Example 2. The case we discuss here is for $c = 0.25$. The results are presented in Fig. 2. The notation $y_{\max}$ is used for the y -coordinate of the maximum of the solution. For $t \leq 8$, the soliton not only moves with a speed, close to $c = 0.25$, but also behaves like a soliton, i.e., preserves its shape, albeit its maximum decreases slightly. For larger times, the solution transforms into a diverging propagating wave, but without a cylindrical symmetry: the fronts are deformed in the direction of propagation. As can be seen from Table 2 the method has second order numerical accuracy in space and time even when the second initial condition is posed at $t = -\tau$ (i.e., (5)₂ is used). This can be attributed to the fact that when $c = 0.25$ the solitary wave tends to preserve its shape, due to the inertia of motion, while for $c = 0$ the tendency towards diverging wave can onset in the very initial moment.

Example 3. In Fig. 3, results for $c = 0.3$ are presented which are second-order accurate in time, similarly to the case $c = 0.25$. For $t < 8$ the behavior of the

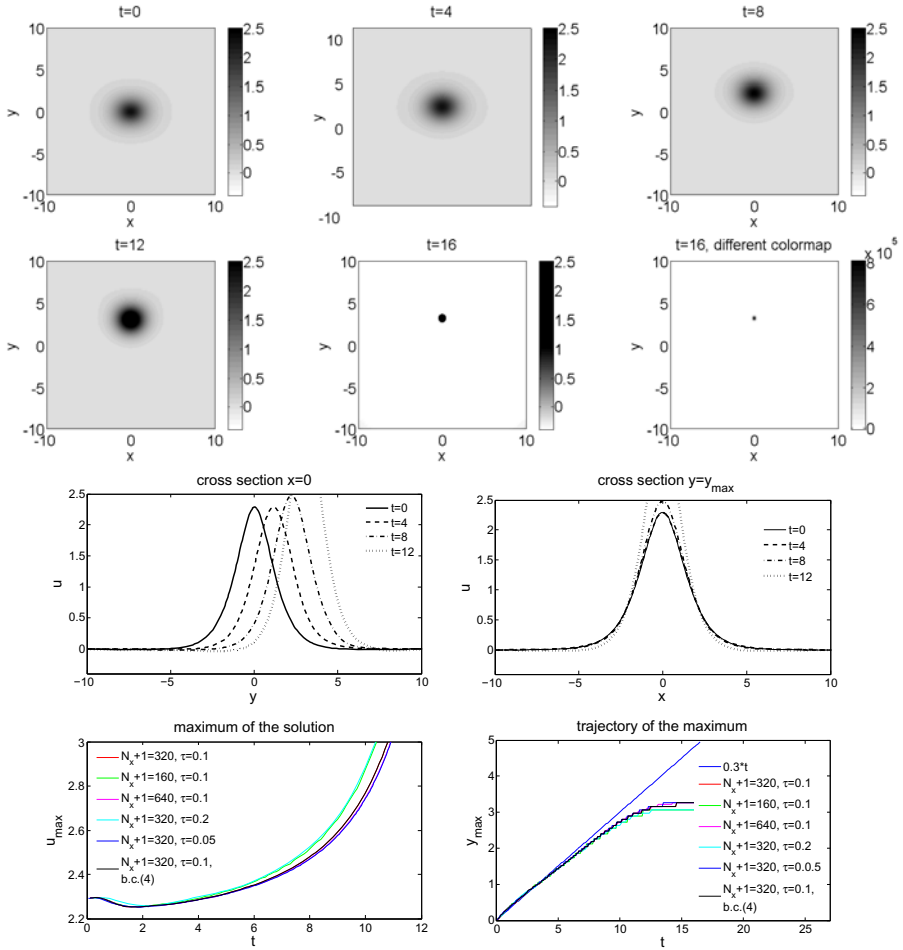


Fig. 3. Evolution of the solution for $c = 0.3$, the evolution of the cross sections at $x = 0$ and $y = y_{\max}$, the maximum $u(0, y_{\max})$, and the trajectory of the maximum

solution is similar to that in the previous example, but for larger times it turns to grow and blows-up for $t \approx 16$. The blow up is connected with the fact that the energy functional is not positive definite for BPE with quadratic nonlinearity (see [10] and the literature cited therein). A threshold value $c = 0.3$ was the last one for which a non-blowing-up evolution was found in [8] on the coarsest grid, while blow-up was encountered on the finest grid. Here we observe blow-up on all grids. This is probably due to the different numerical method used.

Example 4. Taking advantage of the efficiency of the algorithm presented here, we have taken the first sight into the interaction of two structures for different values of their phase speeds. The results are only preliminary, but they are important for answering the question of whether the stationary propagating

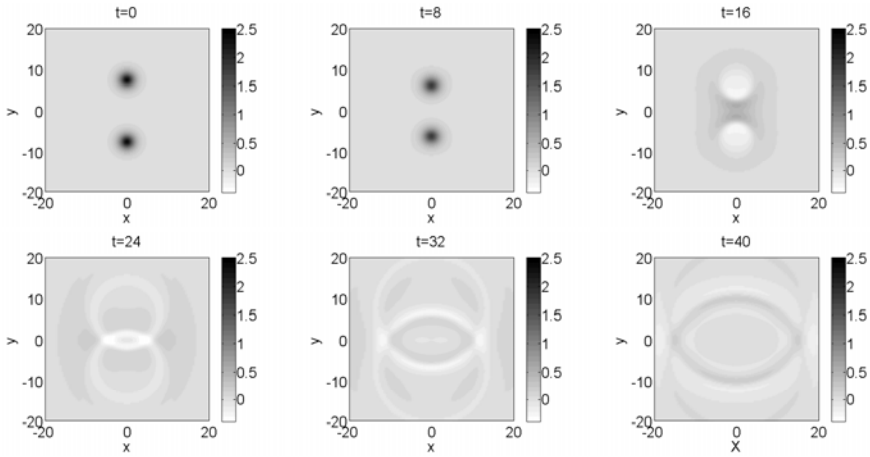


Fig. 4. Evolution of two interacting structures for $c = 0.15$

shapes are actually solitons if they are allowed to interact. In most of the cases with $c_1 = -c_2 \geq 0.2$ (and various initial distances between the structures), the solution blows up after the two structures clash. It is interesting that the threshold for the blow-up is lower than for the evolution of a single structure. We have been able to find non-blowing evolution for $c_1 = -c_2 = 0.15$, only when the initial distance between the centers of the structures is not very small, so the dispersion has some time to begin acting. The result is shown in Fig. 4, where the initial distance is 15. Indeed, the two structures have enough time to set on the track of dispersing waves (concentric diverging circles), and when the latter hit each other, a clear interference pattern onsets. The interaction is similar to the 1D case: they pass through each other. For the largest time $t = 40$ considered, the structures do not seem to have reemerged from the interaction because of their spread, but the centers of the ‘rings’ are well separated. In this sense the 2D structures under investigation can be termed ‘aging coherent structures.’ The detailed investigation of this issues requires a large set of numerical experiments, which goes beyond the frame of the present short note. What is important is that the developed here numerical tool is capable of solving the complex problem at hand.

4 Conclusion

In the present paper, a difference scheme for finding the time dependent localized solutions of the Boussinesq Paradigm Equation (BPE) in two spatial dimensions is devised. The grid is non-uniform and the truncation error is second order in space and time. To reduce the effects connected with the finite size of the computational domain, a special approximation of the asymptotic boundary conditions is used, in which the solution is matched to the expected asymptotic behavior at infinity.

In order to get insight into the possible quasi-particle (solitonic) behavior, results are obtained for the time evolution of supposedly stationary propagating waves for different phase speeds, whose profiles are available from the literature. We have found that for phase speeds $0 \neq c < 0.3$, the initially localized wave disperse in the form of ring-wave expanding to infinity. Respectively, for $c \geq 0.3$ the initial evolution resembles a stationary propagation, but after some period of time a blow-up of the solution takes place. This is in very good quantitative agreement with [8], where a similar (slightly higher threshold) is established for the appearance of the blow-up.

The fact that for $c \approx 0.3$, an time interval exists in which the solution is virtually preserving its shape whilst steadily translating means that 2D solitons could be found in the class of the BPEs. This means that the nonlinearity is strong enough to balance the dispersion which is now much stronger than in the 1D case. In order to firmly establish this fact, our future plans are to consider also equation with different nonlinearity for which the blow-up is not possible.

References

1. Boussinesq, J.V.: Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *Journal de Mathématiques Pures et Appliquées* 17, 55–108 (1872)
2. Zabusky, N.J., Kruskal, M.D.: Interaction of 'solitons' in collisionless plasma and the recurrence of initial states. *Phys. Rev. Lett.* 15, 240–243 (1965)
3. Christov, C.I.: An energy-consistent Galilean-invariant dispersive shallow-water model. *Wave Motion* 34, 161–174 (2001)
4. Christou, M.A., Christov, C.I.: Fourier-Galerkin method for 2D solitons of Boussinesq equation. *Math. Comput. Simul.* 74, 82–92 (2007)
5. Choudhury, J., Christov, C.I.: 2D solitary waves of Boussinesq equation. In: *ISIS Int. Symp. Interdisc. Sci., Natchitoches 2004, APS Conf. Proc.*, vol. 755, pp. 85–90 (2005)
6. Christov, C.I., Choudhury, J.: Perturbation solution for the 2D shallow-water waves. *Mech. Res. Commun.* (submitted)
7. Christov, C.I.: Numerical implementation of the asymptotic boundary conditions for steadily propagating 2d solitons of Boussinesq type equations. *Math. Comp. Simulat.* (accepted)
8. Chertock, A., Christov, C.I., Kurganov, A.: Central-upwind schemes for the Boussinesq paradigm equation. In: *Proc. 4th Russian-German Advanced Research Workshop on Computational Science and High Performance Computing* (2010) (accepted)
9. Kolkovska, N.: Two Families of Finite Difference Schemes for Multidimensional Boussinesq Equation. In: *AIP Conference Series* (accepted)
10. Christov, C.I., Velarde, M.G.: Inelastic interaction of Boussinesq solitons. *J. Bifurcation & Chaos* 4, 1095–1112 (1994)
11. van der Vorst, H.: Iterative Krylov methods for large linear systems. *Cambridge Monographs on Appl. and Comp. Math.* 13 (2009)

Numerical Investigation of Spiral Structure Solutions of a Nonlinear Elliptic Problem

Milena Dimova¹ and Stefka Dimova²

¹ Institute of Mathematics and Informatics, Bulgarian Acad. Sci.,
Acad. G. Bonchev Str., bl. 8, 1113 Sofia, Bulgaria

`mkoleva@math.bas.bg`

² Faculty of Mathematics and Informatics, University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria

`dimova@fmi.uni-sofia.bg`

Abstract. The nonlinear elliptic problem considered arises when investigating a class of self-similar solutions of a reaction-diffusion equation. We focus our study on the solutions of spiral structure. The proposed approach is based on the continuous analog of the Newton's method and on the Galerkin finite element method. To reveal solutions of spiral structure appropriate initial approximations are used. The last ones are expressed by the confluent hypergeometric function ${}_1F_1(a, b; z)$. Algorithms for accurate, fast and reliable computation of its values for broad ranges of the parameters a and b and of the variable z are worked out. A detailed numerical analysis of the evolution of the spiral structure solutions with respect to the medium parameters, including critical values, is carried out.

1 Introduction

A wide variety of spiral patterns can be observed in the physical world - from the tiny twisted biological molecules through the nautilus and ammonites to the curling arms of many galaxies. The spirals play an important role in the growth processes of many biological forms and organisms. Thus it is not surprising that researchers from various fields of science are interested in identifying these patterns and defining them in scientific terms. But it might be surprising, that a family of spiral structures can be described by the solutions of a single 2D nonlinear reaction-diffusion equation of real coefficients - the well known mathematical model of the heat structures [2], introduced and widely investigated by the Russian school of the mathematicians Samarskii and Kurdyumov.

The 2D mathematical model of the heat structures in polar coordinates reads:

$$u_t = \frac{1}{r} \frac{\partial}{\partial r} \left(r u^\sigma \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \varphi} \left(u^\sigma \frac{\partial u}{\partial \varphi} \right) + u^\beta, \quad (1)$$
$$t > 0, \quad 0 < r < \infty, \quad 0 \leq \varphi < 2\pi,$$

where $u(r, \varphi) \geq 0$ is the temperature, the heat conductivity coefficient u^σ and the selfgenerating volume source u^β are functions of the temperature, $\sigma > 0$ and $\beta > 1$ are medium parameters.

Blow-up self-similar solutions of the kind

$$\begin{aligned} u(t, r, \varphi) &= g(t)\theta(\xi, \phi), \quad g(t) = (1 - t/T_0)^{-1/(\beta-1)}, \\ \xi &= r(1 - t/T_0)^{-m}, \quad m = \frac{\beta - \sigma - 1}{2(\beta - 1)}, \quad \phi = \varphi + \frac{C_0}{\beta - 1} \ln(1 - t/T_0) \end{aligned} \quad (2)$$

are found by using the method of invariant-group analysis [5]. Here $T_0 > 0$ is the blow-up time, C_0 is a parameter of the family of solutions. For $C_0 \neq 0$ from (2) it follows

$$r(t)e^{s\varphi(t)} = r(0)e^{s\varphi(0)} = \xi e^{s\phi} = \text{const}, \quad s = (\beta - \sigma - 1)/(2C_0). \quad (3)$$

The dependence (3) means that the trajectories of the inhomogeneities in the medium (for example, local maxima) would be logarithmic spirals for $\beta \neq \sigma + 1$ or circles for $\beta = \sigma + 1$. For $C_0 > 0$ the direction of moving of the inhomogeneities is from the center along the spiral when $\beta < \sigma + 1$ and towards the center when $\beta > \sigma + 1$. The function $\theta(\xi, \phi) \geq 0$ defines the space-time structure of the self-similar solution (2). We put (2) in equation (1), set $T_0 = (\beta - 1)^{-1}$ for convenience, and come to the following nonlinear elliptic equation

$$\begin{aligned} \mathcal{L}(\theta) &\equiv -\frac{1}{\xi} \frac{\partial}{\partial \xi} \left(\xi \theta^\sigma \frac{\partial \theta}{\partial \xi} \right) - \frac{1}{\xi^2} \frac{\partial}{\partial \phi} \left(\theta^\sigma \frac{\partial \theta}{\partial \phi} \right) + \frac{\beta - \sigma - 1}{2} \xi \frac{\partial \theta}{\partial \xi} - C_0 \frac{\partial \theta}{\partial \phi} \\ &+ \theta - \theta^\beta = 0, \quad 0 < \xi < \infty, \quad 0 \leq \phi < 2\pi. \end{aligned} \quad (4)$$

Thus equation (4) has two homogeneous solutions: $\theta_H^0 \equiv 0$ and $\theta_H^1 \equiv 1$.

The case $C_0 = 0$ is widely analyzed [9,7,8]. Radially nonsymmetric solutions of complex symmetry vanishing at infinity are found and investigated for $\beta > \sigma + 1$. For $\beta \leq \sigma + 1$ only simple radially symmetric solutions vanishing at infinity are known to exist [12].

The idea to seek for solutions of equation (4) tending at infinity to the nontrivial constant solution $\theta_H^1 \equiv 1$ was crucial for finding complex symmetry ($C_0 = 0$) and spiral symmetry solutions ($C_0 \neq 0$) for $\beta < \sigma + 1$. This idea was sustained by the radially symmetric case, $\beta < \sigma + 1$. It was shown in [12], that a continuum set of solutions, tending to the nonzero constant solution exists. They oscillate around θ_H^1 and the oscillations are damped. For equation (4) solutions with similar behaviour were firstly numerically constructed in [3].

The *main goal* of this paper is to report results of numerical investigation of the spiral structure solutions ($C_0 \neq 0$) of equation (4) and their dependence on the parameters C_0 , σ and $1 < \beta < \sigma + 1$, including the critical values: $\beta \rightarrow \sigma + 1 - 0$ and $\beta \rightarrow 1 + 0$.

There are two crucial points of the numerical realization of this goal. The first one is to find a more precise than in [3] boundary condition at $\xi = l \gg 1$. The second one is to work out an accurate, fast and reliable computation of the initial approximations to the different solutions of equation (4) for a given set of parameters.

In the next section the initial approximations are introduced and a more precise boundary condition is derived. In section 3 the numerical methods used to solve the nonlinear self-similar problem are briefly described and their accuracy is verified. Section 4 contains methods for computing the confluent hypergeometric function. The results of the parametric investigation of the spiral structures are shown and discussed in the last section. Some open problems are posed as well.

2 Initial Approximations

Using the assumption for small oscillations of the solution $\theta(\xi, \phi)$ around the homogeneous background $\theta_H^1 \equiv 1$, i. e.,

$$\theta(\xi, \phi) = 1 + \alpha y(\xi, \phi), \quad \alpha = \text{const}, \quad |\alpha y| \ll 1$$

and the idea of linearization around it [12], the following linear equation for $y(\xi, \phi)$ is found [4]:

$$-\frac{1}{\xi} \frac{\partial}{\partial \xi} \left(\xi \frac{\partial y}{\partial \xi} \right) - \frac{1}{\xi^2} \frac{\partial^2 y}{\partial \phi^2} + \frac{\beta - \sigma - 1}{2} \xi \frac{\partial y}{\partial \xi} - C_0 \frac{\partial y}{\partial \phi} + (1 - \beta)y = 0.$$

For $\beta \neq \sigma + 1$ particular solutions of the kind

$$y_k(\xi, \phi) = \text{Re}(\xi^k {}_1F_1(a, b; z) e^{ik\phi})$$

which are bounded at $\xi = 0$ are found therein. Here ${}_1F_1(a, b, z)$ is the confluent hypergeometric function, k is a natural number,

$$a = -\frac{\beta - 1}{\beta - \sigma - 1} + \frac{k}{2} - \frac{C_0 k i}{\beta - \sigma - 1}, \quad b = 1 + k, \quad z = \frac{\beta - \sigma - 1}{4} \xi^2.$$

It is shown that it suffices to examine only the case $k > 0$, $C_0 > 0$ and thus the functions $y_k(\xi, \phi)$ are periodic of period $2\pi/k$. The detailed numerical investigation of the functions $y_k(\xi, \phi)$ given in [4] has shown that for large values of ξ they are almost logarithmic spirals, as well as the functions

$$\tilde{\theta}_k(\xi, \phi) = 1 + \alpha y_k(\xi, \phi), \quad |\alpha y_k| \ll 1 \quad (5)$$

are very closed to the sought after solutions $\theta(\xi, \phi)$. These facts gave the possibility first, to find the asymptotics of the solutions of equation (4), and second, to use the functions (5) as initial approximations to the sought after different solutions of (4).

Below we study the case $\beta < \sigma + 1$. Using the asymptotic expansion of ${}_1F_1(a, b, z)$ for $|z| \rightarrow \infty$ [1]

$${}_1F_1(a, b; z) \sim \frac{\Gamma(b)}{\Gamma(b-a)} (-z)^{-a}, \quad \text{Re}(z) \rightarrow -\infty$$

we get

$$y_k(\xi, \phi) \sim \operatorname{Re}(c\xi^{1/m} e^{i(\frac{k}{s} \ln \xi + k\phi)}) \sim |c|\xi^{1/m} \cos(k\phi + \frac{k}{s} \ln \xi + \mu), \quad \xi \rightarrow \infty,$$

where $c = \frac{\Gamma(b)}{\Gamma(a-b)} \left(\frac{\sigma+1-\beta}{4} \right)^{-a}$, $\mu = \arccos \frac{\operatorname{Re}(c)}{|c|}$. The above asymptotic expression predicts the following more precise asymptotics for $\theta(\xi, \phi) = \theta_k(\xi, \phi)$, $k = 1, 2, \dots$, $\xi \rightarrow \infty$:

$$\theta_k(\xi, \phi) \sim 1 + \gamma|c|\xi^{1/m} \cos(k\phi + \frac{k}{s} \ln \xi + \mu), \quad \gamma = \operatorname{const}, \quad |\gamma|c|\xi^{1/m}| \ll 1. \quad (6)$$

Using the asymptotics (6) a boundary condition at $\xi = l \gg 1$ can be deduced to close the self-similar problem:

$$\begin{aligned} \mathcal{L}(\theta_k) \equiv & -\frac{1}{\xi} \frac{\partial}{\partial \xi} \left(\xi \theta_k^\sigma \frac{\partial \theta_k}{\partial \xi} \right) - \frac{1}{\xi^2} \frac{\partial}{\partial \phi} \left(\theta_k^\sigma \frac{\partial \theta_k}{\partial \phi} \right) + \frac{\beta - \sigma - 1}{2} \xi \frac{\partial \theta_k}{\partial \xi} - C_0 \frac{\partial \theta_k}{\partial \phi} \\ & + \theta_k - \theta_k^\beta = 0, \quad 0 < \xi < l, \quad 0 \leq \phi < 2\pi/k, \end{aligned} \quad (7)$$

$$\lim_{\xi \rightarrow 0} \xi \theta_k^\sigma \frac{\partial \theta_k}{\partial \xi} = 0, \quad \phi \in [0, 2\pi/k],$$

$$\frac{\partial \theta_k}{\partial \xi} = \frac{\theta_k - 1}{m\xi} - \frac{\gamma|c|k}{s\xi^{(m-1)/m}} \sin(k\phi + \frac{k}{s} \ln \xi + \mu), \quad \xi = l \gg 1, \quad \phi \in [0, 2\pi/k], \quad (8)$$

$$\theta_k(\xi, 0) = \theta_k(\xi, 2\pi/k), \quad \frac{\partial \theta_k}{\partial \phi}(\xi, 0) = \frac{\partial \theta_k}{\partial \phi}(\xi, 2\pi/k), \quad 0 \leq \xi \leq l.$$

3 Numerical Method for the Self-similar Problem

The method is presented in detail in our previous work [3]. Here only the main steps are briefly described.

To solve the nonlinear boundary value problem (7), (8) an iterative algorithm based on the continuous analog of Newton's method (CANM) [11] is used.

When applied to the nonlinear equation $\mathcal{L}(\theta) = 0$, the CANM leads to the iteration process

$$\mathcal{L}'(\theta_n)v_n = -\mathcal{L}(\theta_n), \quad (9)$$

$$\theta_{n+1} = \theta_n + \tau_n v_n, \quad 0 < \tau_n \leq 1, \quad n = 0, 1, \dots, \quad (10)$$

$$\theta_0 = \tilde{\theta}_k(\xi, \phi). \quad (11)$$

Here $\mathcal{L}'(\theta_n)$ is the Frechét derivative of the operator $\mathcal{L}$ at the point θ_n ; θ_0 is the initial approximation (5) to one of the sought after different solutions $\theta_k(\xi, \phi)$ for given parameters σ, β . For convenience the subscript k is omitted in (9), (10).

The equation (9) is linear with respect to the iteration corrections v_n . To solve it, we use the Galerkin finite element method and bilinear elements. At each step

of the iteration process (9) – (11) we get a linear algebraic system of equations $AV = B$ with nonsymmetric matrix. It is stored and used in sky-line form. The linear algebraic problems are solved by using LU-decomposition.

The accuracy of the described methods is experimentally analyzed using embedded grids. Table 1 shows the values of the spiral structure solution computed for parameters $\sigma = 2, \beta = 2.4, C_0 = 1, k = 1, \xi \in [0, 14], \phi \in [0, 2\pi]$ at some common points of embedded grids $h, h/2, h/4, h = (h_\xi, h_\phi)$. The order of accuracy α is computed by Runge's method $\alpha = \ln |(\theta^h - \theta^{h/2})/(\theta^{h/2} - \theta^{h/4})| \ln^{-1} 2 \approx 2$.

Table 1. Spiral structure solution for $\sigma = 2, \beta = 2.4, k = 1, C_0 = 1$

h_ξ	h_ϕ	$\theta(0, 0)$	$\theta(1.6, 0.418879)$	$\theta(10, 3.141593)$	$\theta(6.6, 5.026548)$
0.2	$\pi/15$	1.0000015920889	1.0002797451706	0.99999054430131	1.0000776972682
0.1	$\pi/30$	1.0000015999898	1.0002804409990	0.99999052381423	1.0000779446091
0.05	$\pi/60$	1.0000016667206	1.0002832580659	0.99999044176990	1.0000789468270
α		3.09	2.02	2.00	2.02

4 Computation of the Initial Approximations

To compute the initial approximations (5) one needs an accurate, fast and reliable computation of the confluent hypergeometric function ${}_1F_1(a, b; z)$ for different parameter regimes within the complex plane for the parameters a and b , as well as for different regimes of the variable z . But this is an extremely difficult task in practice. The reason of this is that the non-trivial structure of the series expansion of ${}_1F_1(a, b; z)$ creates many numerical issues such as cancelation and round-off error, as well as the existence of very large alternating terms, which become especially significant for certain ranges of the parameters and the variable. The goal is to choose appropriate methods for the different ranges of a, b and z .

Let's consider how the values of a and z change when σ and β vary. For fixed values of σ ($1 < \sigma < 6$ for some real-life problems) there are two critical values for β . The first one is $\beta \rightarrow \sigma + 1 - 0$, when $\text{Re}(a)$ and $\text{Im}(a)$ increase extremely fast. The second one is $\beta \rightarrow 1 + 0$, when a gets moderate values, but the asymptotics of the initial approximations decays very slowly when ξ increases (see (6)), so we need really large computational interval for z . That is why we suggest the following algorithm for computing ${}_1F_1(a, b; z)$.

4.1 Taylor Series Expansion

For moderate values of $|a|$ and $|z|$ ($|a| < 50, |z| < 100$) we use the Taylor series expansion. Because of its very large alternating terms first we apply the transformation

$${}_1F_1(a, b; z) = e^z {}_1F_1(b - a, b; -z) = e^z {}_1F_1(p, q; w).$$

Then we implement the basic power series definition:

$${}_1F_1(p, q; w) = \sum_{j=0}^{\infty} \frac{(p)_j}{(q)_j} \frac{1}{j!} w^j = \sum_{j=0}^{\infty} A_j w^j.$$

The computation can be carried out using the following procedure:

$$A_0 = 1, S_0 = A_0, A_{j+1} = A_j \times \frac{(p+j)}{(q+j)} \times \frac{w}{(j+1)}, S_{j+1} = S_j + A_{j+1}, j = 0, 1, 2, \dots$$

The stopping criterion we use is $\frac{|A_{N+1}|}{|S_N|} < tol = 10^{-15}$ and $\frac{|A_N|}{|S_{N-1}|} < tol$. This method produces accurate and fast results with up to 14-15 digits of accuracy.

4.2 Asymptotic Series

The above proposed method is not applicable for large values of $|z|$ (typically the methods cease to be effective for $|z| > 100$). In such a case we use the asymptotic expansion for $|z| \rightarrow \infty$, $z \in \mathbb{R}$ [1]:

$$\begin{aligned} {}_1F_1(a, b; z) &= \frac{\Gamma(b)e^{i\pi a}z^{-a}}{\Gamma(b-a)} \sum_{j=0}^{\infty} \frac{(a)_j(1+a-b)_j}{j!} (-z)^{-j} \\ &+ \frac{\Gamma(b)e^z z^{a-b}}{\Gamma(a)} \sum_{j=0}^{\infty} \frac{(b-a)_j(1-a)_j}{j!} (z)^{-j}. \end{aligned}$$

In our computations we use the same techniques as for the Taylor series method. To compute the Gamma function we use the effective code based on the ideas from [13] for complex argument.

4.3 Expansion in Ascending Series of Chebyshev Polynomials

Frequently, the robustness of a method for computing the confluent hypergeometric function is greatly reduced by its poor performance as $|Re(a)|$ gets larger. The recurrence relation techniques [6] can reduce the problem to a simpler problem of computing ${}_1F_1(a, b; z)$ for values of $|Re(a)|$ closer to 0. This method is not applicable in our case because both the real and the imaginary part of a increase as $\beta \rightarrow \sigma + 1 - 0$. That is why for large value of $|a|$ ($|a| > 50$) we use the expansion in ascending series of Chebyshev polynomials [10]:

$${}_1F_1(a, b; z) = \sum_{n=0}^{\infty} C_n(w) T_n^*(z/w), \quad 0 \leq z/w \leq 1,$$

where $T_n^*(x)$ are the shifted Chebyshev polynomials of the first kind, and the coefficients $C_n(w)$ satisfy the recurrence formula

$$\begin{aligned} \frac{2C_n}{\varepsilon_n} &= \frac{n+1}{(n+a)(n+2)} \left[\frac{4(n+b)(n+2)}{w} - (n+3-a) \right] C_{n+1} \\ &+ \left[1 + \frac{4(n+3-b)(n+1)}{n+a} \right] C_{n+2} + \frac{(n+1)(n+3-a)}{(n+a)(n+2)} C_{n+3}, \end{aligned}$$

$\varepsilon_0 = 1$ and $\varepsilon_n = 2$ for $n > 0$. The coefficients $C_n(w)$ can be found by use of the recursion formula in backward direction, together with the normalization relation $\sum_{n=0}^{\infty} (-1)^n C_n(w) = 1$.

5 Parametric Investigation

We have investigated the evolution of the spiral structure solutions of problem (7), (8) depending on the parameters k , C_0 , σ , β .

The parameter k determines the number of the spiral arms. On Fig. 1 the graphs of the solutions for $k = 1$ (one-armed spiral), $k = 2$ (two-armed spiral), $k = 3$ (three-armed spiral) are shown. The rest of parameters are $\sigma = 3$, $\beta = 3.6$, $C_0 = 1$. For fixed other parameters, C_0 determines the density of the spirals. Fig. 2 demonstrates the changing of the spiral density when C_0 takes values $C_0 = 1, 2, 3$. The increasing of the core of the spiral when $\beta \rightarrow 1 + 0$ and $\sigma = 3$, $k = 1$, $C_0 = 1$ is shown on Fig. 3. The core of a spiral is the circle of radius ξ_0 around the origin, out of which $|\theta(\xi, \phi) - 1| < 0.01 \max_{\xi, \phi} |\theta(\xi, \phi) - 1|$, $\forall \xi > \xi_0$, $\forall \phi \in [0, 2\pi/k]$. When β decreases from $\beta = 2.9$ to $\beta = 1.5$, the computational interval for ξ increases from $[0, 14]$ to $[0, 5000]$. On Fig. 4 the spiral structure solutions for $\sigma = 3$, $k = 1$, $C_0 = 1$ and $\beta = 3.1, 3.4, 3.7, 3.9, 3.92, 3.96$ are given. The number of turnings of the spirals increases with β and the logarithmic spirals approach Archimedean ones.

Let us mention, there is no theoretical investigations about the number of the different solutions of problem (7), (8) for fixed values of the parameters σ , β , C_0 , k . For different values of the constant γ in the boundary condition (8) (see also (6)) we obtain solutions of the same structure but of different amplitudes (deviations from θ_H^1). The ranges of the parameters C_0 , k , where the solutions exist, are not known as well.

The most challenging questions concern the existence of spiral structure solutions for $\beta \geq \sigma + 1$. The main difficulty in this case is to find appropriate conditions for $\xi \rightarrow 0$ and $\xi \rightarrow \infty$.

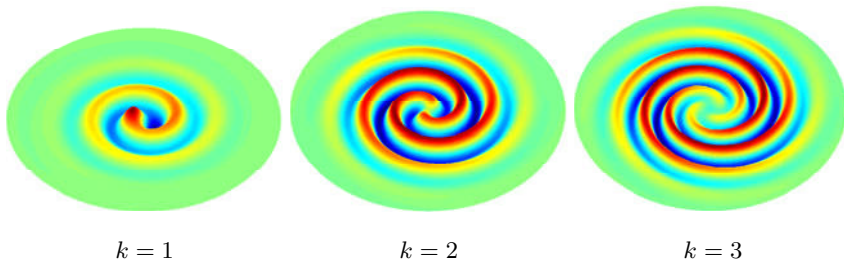


Fig. 1. One-armed spiral solution ($k = 1$), two-armed spiral solution ($k = 2$), three-armed spiral solution ($k = 3$), $C_0 = 1$, $\sigma = 3$, $\beta = 3.6$

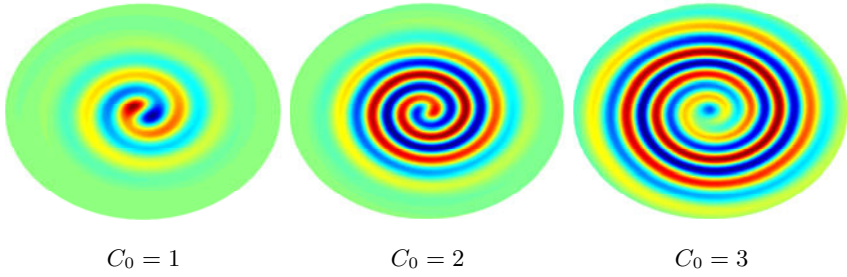


Fig. 2. One-armed spiral solution for various values of C_0 : $\sigma = 3$, $\beta = 3.6$, $k = 1$, $C_0 = 1, 2, 3$

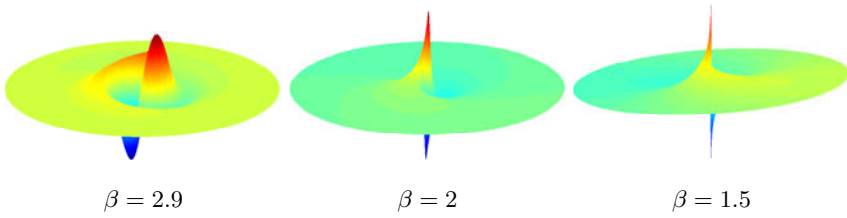


Fig. 3. Evolution of a one-armed spiral solution depending on the β : $\sigma = 3$, $C_0 = 1$, $k = 1$, $\beta = 2.9, 2, 1.5$

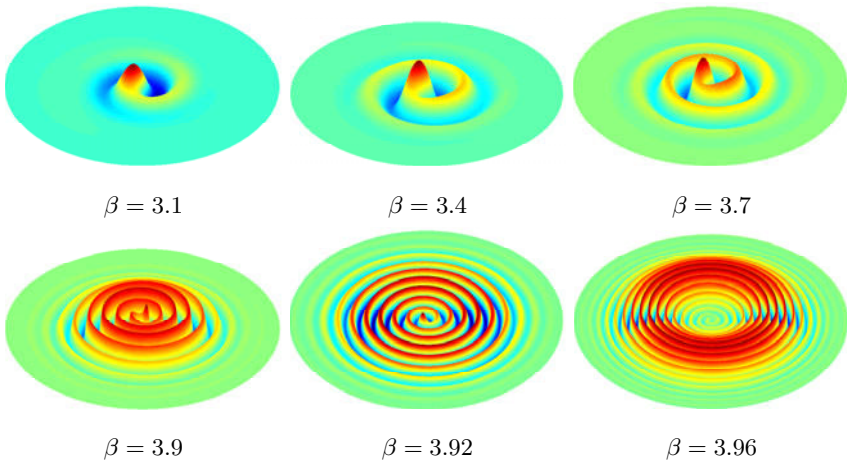


Fig. 4. Evolution of a one-armed spiral solution depending on the β : $\sigma = 3$, $C_0 = 1$, $k = 1$, $\beta = 3.1, 3.4, 3.7, 3.9, 3.92, 3.96$

Acknowledgments. This work is partially supported by Sofia University Scientific foundation under Grant No 196/2010.

References

1. Abramovitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards (1970)
2. Akhromeeva, T.S., Kurdyumov, S.P., Malinetskii, G.G., Samarskii, A.A.: Chaos and Dissipative Structures in Reaction-Diffusion Systems. Nauka, Moscow (1992)
3. Dimova, S.N., Kastchiev, M.S., Koleva, M.G., Vasileva, D.P.: Numerical Analysis of Radially Nonsymmetric Blow-up Solutions of a Nonlinear Parabolic Problem. *J. Comp. Appl. Math.* 97, 81–97 (1998)
4. Dimova, S.N., Vasileva, D.P.: Numerical Realization of Blow-up Spiral Wave Solutions of a Nonlinear Heat-Transfer Equation. *Int. J. Num. Meth. Heat Fluid Flow* 4, 497–511 (1994)
5. Galaktionov, V.A., Dorodnicyn, V.A., Elenin, G.G., Kurdyumov, S.P., Samarskii, A.A.: The Quasilinear Heat Conduction Equation with a Source: Enhancement, Localization, Symmetry, Exact Solutions, Asymptotic Forms and Structures. *J. Sov. Math (JOSMAR)* 41, 1163–1356 (1988)
6. Gil, A., Segura, J., Temme, N.M.: Numerical Methods for Special Functions. SIAM, Philadelphia (2007)
7. Koleva, M.G., Dimova, S.N., Kaschiev, M.S.: Analysis of the Eigen Functions of Combustion of a Nonlinear Medium in Polar Coordinates. *Math. Modeling* 3, 76–83 (1992)
8. Kurkina, E.S., Nikol'ski, I.M.: Bifurcation Analysis of the Spectrum of Two-Dimensional Thermal Structures Evolving with Blow-up. *Comp. Math. and Modeling* 17(4), 320–340 (2006)
9. Kurdyumov, S.P., Kurkina, E.S., Potapov, A.B., Samarskii, A.A.: Complex Multi-dimensional Structures of Combustion of a Nonlinear Medium. *Dokl. Acad. Nauk SSSR* 274, 1071–1075 (1984)
10. Luke, Y.: Algorithms for the Computation of Mathematical Functions. Academic Press, London (1977)
11. Puzynin, I.V., et al.: Methods of Computational Physics for Investigation of Models of Complex Physical Systems. *Particals & Nucley* 38 (2007)
12. Samarskii, A.A., Galaktionov, V.A., Kurdyumov, S.P., Mikhailov, A.P.: Blowup in Problems for Quasilinear Parabolic Equations. Walter de Gruyter, Berlin (1988)
13. Zhang, S., Jin, J.: Computation of Special Functions. John Wiley & Sons, Chichester (1996)

Bidirectional Beam Propagation Method Applied for Lasers with Multilayer Active Medium

N.N. Elkin, A.P. Napartovich, and D.V. Vysotsky

State Science Center Troitsk Institute for Innovation and Fusion Research(TRINITI),
142190, Troitsk Moscow Region, Russia
`elkin@triniti.ru`

Abstract. The vertical external cavity surface emitting laser (VECSEL) as a typical example of laser with multilayer active medium is considered. The round-trip operator technique is presented in the given paper based on the bidirectional beam propagation method (BiBPM). Similarly to traditional Fox-Li technique our method not requires explicit calculation of matrix of the round-trip operator and suits perfectly to Krylov subspace methods of linear algebra. The presented method is extended in natural way to non-linear case taking into account light-medium interaction. The results of modeling of a VECSEL with a resonant array of quantum wells are presented.

1 Introduction

Optical devices that have piecewise continuous gain and index distributions along the main propagation direction are widespread. A resonant heterostructure of an array of quantum wells (QW) is of practical interest for application in VECSELs. The steady-state oscillating modes of a laser are described by non-linear partial differential equations containing eigenvalues. Book [1] can be recommended as the general work on the solution of nonlinear eigenvalue problems. However, it should be noted that the theory of nonlinear eigenvalue problems is far from completion. The multilayer medium in the laser cavity complicates considerably the mathematical modeling because of partial reflections from the layer interfaces. For the first time, the applications of BiBPM for laser devices [2] were restricted by linear eigenvalue problem neglecting influence of the light beam on gain and index of the active medium. The eigenvalue problems for a non-hermitian matrix of high dimension were solved numerically in [2]. Next, the BiBPM combined with the round-trip operator technique was developed for self-consistent solution of wave field and material equations [3]. The Krylov subspace methods [4] applied in [3] to calculate the eigenfunctions of the linear wave equation is considerably more effective in comparison with the matrix method [2].

In present paper the modification of the algorithm [3] was applied for a VECSEL with a resonant array of quantum wells. Modeling of a VECSEL using diffraction theory approximation and taking into account diffusion equations for charge carriers in QWs was performed for the first time to our knowledge.

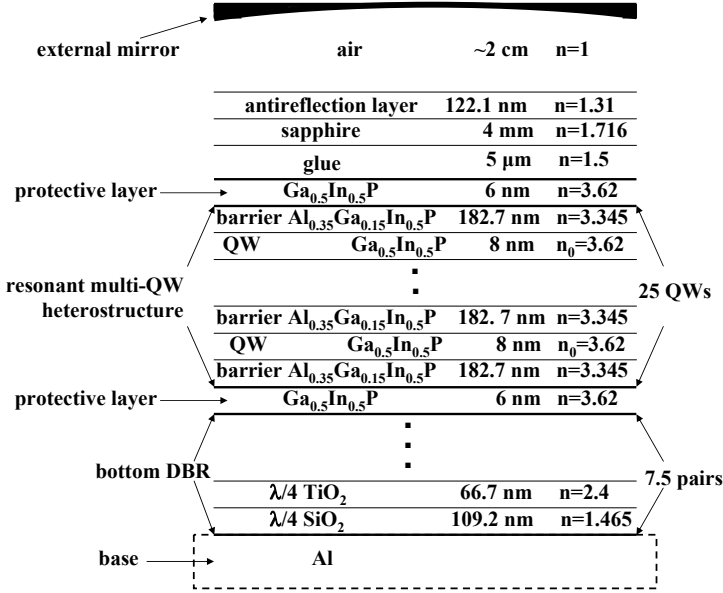


Fig. 1. Scheme of the VECSEL (Cross-Section View)

2 Description of the Device and Basic Equations

The scheme of the VECSEL containing a resonant heterostructure is presented in Fig. 1. Assuming a vertical z -axis we represent the VECSEL as a pile of layers, separated by planes $\{z = z_k, k = 0, \dots, M\}$ where M is the total number of layers. The index and absorption are constant in each layer except the active layers (QWs) where non-uniform distributions are controlled by electrical current and light intensity. To distinguish QWs from other layers we define the index array $\{\nu(l), l = 1, \dots, q\}$, where $q = 25$ is the number of QWs. If $k = \nu(l)$ then the layer $[z_{k-1}, z_k]$ is the l -th QW. The spherical external mirror has radius of curvature 3 cm. Optical length L_{opt} of the space between mirror and heterostructure is a variable parameter.

We assume that the scalar diffraction theory is applicable. The pump profile has circular symmetry according to assumption. Therefore, we use cylindrical coordinates. Laser modes have a time dependence of the form $E(r, \varphi, z, t) = U(r, \varphi, z) \exp(-i\Omega t)$, $\Omega = \omega_0 + \Delta\omega - i\delta$, where ω_0 is the reference frequency, $\Delta\omega = \omega - \omega_0$ is the frequency shift and δ is the attenuation factor. The reference wavenumber and reference wavelength are defined by standard relations: $\omega_0 = k_0 c$, $k_0 = 2\pi/\lambda_0$. The solutions of a form $U(r, \varphi, z) = U_m(r, z) \exp(im\varphi)$ are subjects of investigation. Introducing new variables $g_t = 2\delta/c$, $\Delta k = \Delta\omega/c$, $\beta = g_t + i2\Delta k$, we have obtained the equation for m -th angular harmonic:

$$\frac{\partial^2 U_m}{\partial z^2} + Q^2 U_m = 0, \quad Q = \sqrt{k_0^2 n^2 - ik_0 g - ik_0 n^2 \beta + \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} - \frac{m^2}{r^2}}. \quad (1)$$

The equation (1) contains a complex eigenvalue β . The real part of β is the decay rate of wave field expressed in units of inverse length, the imaginary part is the twice wavenumber shift relative to the reference value. Here n and g are index and gain respectively, Q is the operator of longitudinal wavenumber. This unusual form of the Helmholtz equation was chosen because vertical direction of wave propagation predominates in VECSEL. The boundary conditions at the interfaces between adjoining layers and at the lateral boundary were determined. We use condition of continuity for the wave field U_m and its normal derivative at the interfaces. Luckily there is no problem of boundary conditions at the lateral boundary because the active layers have strong attenuation in the absence of pump. The boundary condition at the mirror suits a good reflecting surface. The set of quantum wells forms a finite periodical structure so as optical length of an one period is equal to $\lambda_s = 640$ nm.

The main problem consists of self-consistent solving of the wave field equation and material equations in order to find the spatial profile of a laser electromagnetic field and its frequency in steady-state mode of operation. We restrict our consideration with axisymmetric laser modes. According to this condition, we are to solve the axisymmetric ($m = 0$) equation (1) jointly with the set of non-linear diffusion equations [5]

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial Y_l}{\partial r} \right) - \frac{Y_l}{D\tau_r} - \frac{B}{D} N_t Y_l^2 - \frac{|U_0|^2 \ln(\chi(Y_l))}{D\tau_r} = \frac{-kE_e j}{DN_t 3E_g qed}, l = 1, \dots, q \quad (2)$$

for normalized carrier density $Y_l = N_l/N_t$ at the l -th active layers. Here N_l is the carrier density, D is the diffusion coefficient, τ_r is a recombination time, B is a coefficient of nonlinearity, d is thickness of the QW, e is the elementary charge, $N_t = \left(-1/\tau_r + \sqrt{1/\tau_r^2 + (4BkE_e j_t)/(3E_g qed)} \right) / (2B)$ is the carrier density for conditions of transparency, j_t is the injection current density for conditions of transparency, $|U_0|^2$ is the normalized light intensity, E_e is the energy of electrons, k is the part of the energy of electrons inputed into QWs, E_g is band gap of the barrier layers, $j = If(r/r_0) (2\pi \int f(r/r_0) r dr)^{-1}$ is the current density of the electron beam (e-beam), I is the total current of the beam, $f(\rho)$ is the pump profile function, r_0 is the pump region radius. Zero boundary conditions for $Y_j(r)$ are set at the lateral boundary of the active layer. The function $\chi(Y)$, gain and index at the active layers are approximated by the formulas

$$\chi(Y) = \begin{cases} \alpha + (1 - \alpha)Y^{1/1-\alpha}, & Y < 1 \\ Y, & Y \geq 1 \end{cases}, \quad (3)$$

$$g_l = g_0 \ln(\chi(Y_l)), \quad n_l = n_0 - R(g_l - g_{\min})/(2k_0),$$

where $\alpha = \exp(g_{\min}/g_0)$, g_0 and $g_{\min}$ are gain parameter, n_0 is the refractive index in the absence of carriers, R is the line enhancement factor. The equation (1) at $m = 0$ jointly with the equations (2) and (3) supplemented with corresponding boundary conditions form the eigenvalue problem for a non-linear operator. The supplementary condition $\delta = 0$ ($\text{Re}(\beta) = 0$) is required for steady-state operation.

We consider also an subsidiary problem when we neglect dependence of material characteristics on electromagnetic field intensity. It is so called case of "frozen" active medium. The equation (1) with the boundary conditions described must be solved in order to find the spatial profile of an eigenfunction and the complex eigenvalue β . The angular-dependent solutions ($m \neq 0$) are considered in this case also.

3 Numerical Solution

According to BiBPM we represent a wave field U in each horizontal plane as a vector $(V^+ V^-)^T$ of the upward and downward propagating waves, so as $U = V^+ + V^-$. The wave fields in two arbitrary planes, marked by symbols t and b are bounded by a transfer equation: $(V_t^+ V_t^-)^T = \mathbf{M} (V_b^+ V_b^-)^T$, where $\mathbf{M}$ is a transfer matrix. Transfer matrix for set of layers can be calculated as a product of the elementary interface and propagation matrices [2]:

$$\mathbf{T}_k = \frac{1}{2} \begin{pmatrix} 1 + Q_{k+1}^{-1} Q_k & 1 - Q_{k+1}^{-1} Q_k \\ 1 - Q_{k+1}^{-1} Q_k & 1 + Q_{k+1}^{-1} Q_k \end{pmatrix}, \quad \mathbf{P}_k = \begin{pmatrix} e^{iQ_k h_k} & 0 \\ 0 & e^{-iQ_k h_k} \end{pmatrix}, \quad (4)$$

where $h_k = z_k - z_{k-1}$, Q_k is the operator of longitudinal wavenumber in the k -th layer. For example, $\mathbf{M} = \mathbf{P}_M \mathbf{T}_M \dots \mathbf{T}_{\nu(q)+1} \mathbf{P}_{\nu(q)+1}$ is the transfer matrix for region above the top QW, $\mathbf{M} = \mathbf{T}_{\nu(l)} \mathbf{P}_{\nu(l)} \mathbf{T}_{\nu(l)-1}$ ($l = 1, \dots, q$) are the transfer matrices for QWs, $\mathbf{M} = \mathbf{P}_{\nu(l)-1}$ ($l = 2, \dots, q$) are the transfer matrices for the barrier layers and $\mathbf{M} = \mathbf{P}_{\nu(1)-1} \mathbf{T}_{\nu(1)-2} \dots \mathbf{P}_1 \mathbf{T}_0$ is the transfer matrix for bottom DBR region. The fast Hankel transform algorithm [6] was used for effective calculations with the transfer matrices. In the wavenumber space the operator Q_k is replaced by the number q_k , the operational matrices $\mathbf{T}_k$ and $\mathbf{P}_k$ became the numerical matrices. The calculations in the QW regions were performed in the physical space because of non-uniform transverse gain and index distributions. The approach of locally uniform wave field was used. This approach is admissible since thickness of the QW is far less then the wavelength.

Joining the set of every possible transfer equations and boundary conditions with the condition of absence of externally injected electromagnetic fields we can obtain the closed system of equations and represent it as an eigenvalue problem:

$$\mathbf{P}(g, n, \beta)u = \gamma u, \quad (5)$$

where $\mathbf{P}$ is the round-trip operator, u is the upward propagating wave at the preselected plane, $\gamma = 1$.

Our approach to problem (5) consists in solution of the auxiliary problem for a function u and eigenvalue γ to be found provided the value of β is specified. The value β is adjusted until $\gamma = 1$ within a certain tolerance. Generally, calculations were organized as follows: inner iteration procedure solves the equation (5) at the fixed value of β to find one or several eigenpairs (u, γ) ; the external iterative cycle encloses the inner cycle and serves to find the value β where $\gamma = 1$.

In case of self-consistent solution the eigenvalue β is an imaginary number, $\beta = i2\Delta k$. The problem for fixed β is the eigenvalue problem for a non-linear operator because gain g and index n are determined by equation (2) and depend on u . This problem is solved by the Fox-Li iteration method [7]. The value Δk is adjusted in an external cycle using the secant method.

For case of "frozen" active medium we have the linear non-hermitian eigenvalue problem if β is fixed. Only several eigenpairs (u, γ) are required. The standard Arnoldi method is efficient in this case. Calculation and storing of matrix of $\mathbf{P}(g, n, \beta)$ is not required. It is necessary to calculate elements of vector $\mathbf{P}(g, n, \beta)u$ only. The complex eigenvalue β is adjusted in an external iteration cycle using the Broyden method [8].

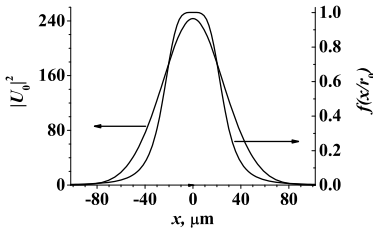


Fig. 2. Cross-section view of the TEM₀₀ lasing mode intensity at the top QW and the pump profile function

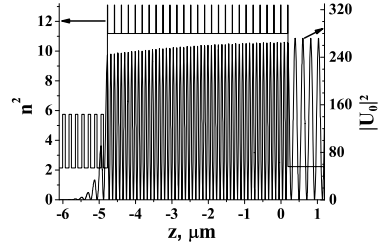


Fig. 3. Axial profiles of n^2 (stepwise function) and light intensity (continuous function)

4 Results and Discussion

The set of QWs forms a resonant heterostructure with resonant wavelength slightly differing from period λ_s . The reference wavelength λ_0 have to be as close as possible to the resonant wavelength with a view to improve accuracy and efficiency of computations. Taking into account the previous work [3], [9] we set the reference wavelength $\lambda_0 = 642.2$ nm. The other parameters were given as follows: $D = 0.5 \text{ cm}^2\text{s}^{-1}$, $\tau_r = 10^{-9}\text{s}$, $B = 3.5 \times 10^{-10}\text{cm}^3\text{s}^{-1}$, $k = 0.75$, $E_e = 4 \times 10^4$ eV, $E_g = 2.36$ eV, $j_t = 2.35 \text{ A cm}^{-2}$, $r_0 = 25 \mu\text{m}$, $g_0 = 3400 \text{ cm}^{-1}$, $g_{\min} = -1000 \text{ cm}^{-1}$, $R = 2.5$, $I = 2.35 \text{ mA}$. Calculations were performed for the profile function $f(\rho) = (1 + \rho^4)^{-1}$. The external spherical mirror has transverse size $400 \mu\text{m}$ and reflection coefficient 0.985. Test calculations reveal that for practical purposes 256 mesh nodes over polar radius r is good choice. The lasing mode intensity in physical units is calculated by the formula $J = J_s|U_0|^2$, where $J_s = (hcN_t)/(\lambda_0 g_0 \tau_r)$ is the intensity of saturation. We use notation TEM_{nm} for optical modes in a VECSEL. Here m is the angular quantum number responding to dependence $\sim \exp(im\varphi)$ and n is the number of mode in case of ascending ordering of decay rate.

In the interval $2.4 \text{ cm} < L_{\text{opt}} < 2.98 \text{ cm}$ we have calculated single-mode operation regimes and found it stable. The last value corresponds to near-concentric

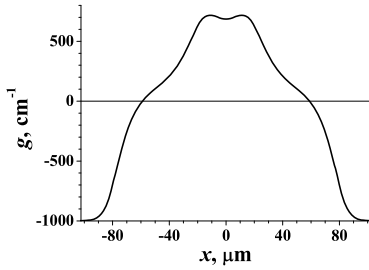


Fig. 4. Cross-section view of gain at the top QW

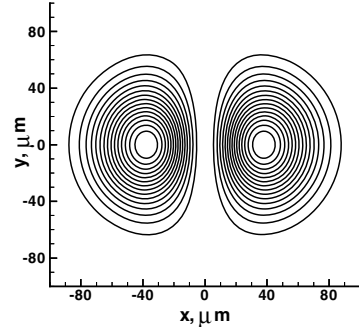


Fig. 5. Contour plot of the intensity of TEM_{01} mode ($g_t = 4.39 \times 10^{-3} \text{ cm}^{-1}$)

configuration of the optical resonator. The results of calculations at $L_{opt} = 2.4 \text{ cm}$ are presented in Figs. 2 - 5. The calculated wavelength is 642.202 nm. The lasing mode intensity has good overlapping with the pump profile as seen in Fig. 2. The longitudinal profile of the lasing mode shown in Fig. 3 has oscillations so as the antinodes of the standing wave are located at the gain layers (QWs). The transverse profile of gain Fig. 4 is distorted due to saturation by the light intensity. The subthreshold TEM_{01} mode intensity distribution at the top QW is shown in Fig. 5. This mode has small decay rate and can destroy single-mode operation under some disturbances. The single-mode operation was not obtained at $L_{opt} = 2.3 \text{ cm}$ because the light intensity tends to zero in iteration process.

To understand this strange effect we have performed calculations for "frozen" active medium formed by e-beam pump ignoring saturation of the active media by the light intensity, i.e. supposing that $|U_0|^2 \equiv 0$ in (2). The dependences of decay rates and wavenumber shifts on L_{opt} are presented in Figs. 6 and 7 for three modes. One can see that in interval $2.5 \text{ cm} < L_{opt} < 2.98 \text{ cm}$ modes are strongly discriminated on losses. It follows from fact that loss caused by diffraction on the mirror

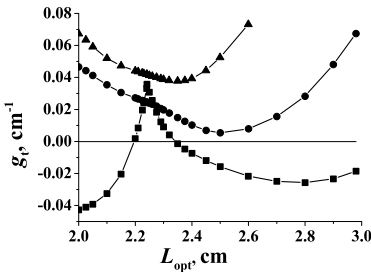


Fig. 6. Decay rate of modes with highest Q -factor: TEM_{00} (squares), TEM_{01} (circles) and TEM_{10} (triangles). Unsaturated medium.

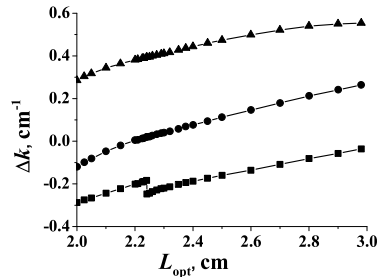


Fig. 7. Wavenumber shift of modes with highest Q -factor: TEM_{00} (squares), TEM_{01} (circles) and TEM_{10} (triangles)

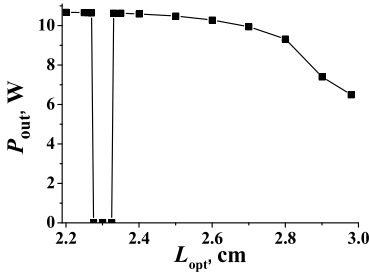
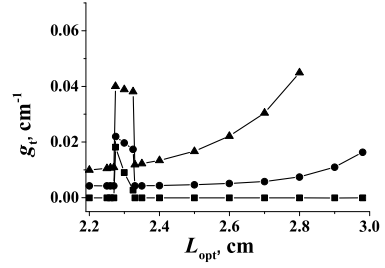


Fig. 8. Output power

Fig. 9. Decay rate of modes with highest Q -factor: operating mode TEM_{00} (squares), TEM_{01} (circles) and TEM_{10} (triangles). Non-linear medium, self-consistent solution.

edge increase at near-concentric configuration. The decay rate of the fundamental mode is considerably less than zero in the specified interval thus the fundamental mode is superthreshold. If $L_{opt} < 2.5$ cm then transverse sizes of modes at the external mirror become lesser then the size of mirror and diffraction losses on the mirror edge become negligible. On the contrary, sizes of modes at the multi-QW structure increase with decreasing L_{opt} and become approximately equal to size of the pump spot as one can see in Fig. 2. The non-uniform gain-index profile formed by e-beam pumping plays a key role in mode profile formation. As a result, mode patterns may remarkably deviate from patterns of classic Laguerre-Gaussian beams. In the interval $2.2 \text{ cm} < L_{opt} < 2.34 \text{ cm}$ all the modes have positive decay rate g_t , and so they are subthreshold. This circumstance explains why we have not obtain laser generation at $L_{opt} = 2.3$ cm. Singular points in Figs. 6 and 7 are the results of change of the fundamental mode.

Finally, we have performed calculations with the framework of self-consistent problem defined by equation (1)-(3) and boundary and supplementary conditions. The output power $P_{out} = 2\pi \int J_{out} r dr$, where $J_{out} = J_s |V_{out}^+|^2$ is the intensity of the outgoing wave, depends smoothly on L_{opt} except for small neighbourhood of the value $L_{opt} = 2.3$ cm where it jumps to zero (Fig. 8). The calculations for "frozen" active medium formed by e-beam pump and the light intensity of the operating mode are presented in Fig. 9. The decay rate $g_t = 0$ for the operating mode (square markers) what confirms steady-state regime of lasing. Other modes have positive decay rate therefore steady-state operation is stable. Exception to the pattern is the interval where $P_{out} = 0$, all the modes have positive decay rate and steady-state operation is impossible.

5 Conclusion

The BiBPM developed for multilayer media can be successfully joined up with the well-known round-trip operator technique for optical resonators including

Fox-Li iterations and Krylov subspace methods. As a result we have developed the efficient numerical method for modeling lasers with multilayer structure including linear and non-linear regimes of operation. The given numerical algorithm allows us to calculate the mode spatial profile, output power, exact wavelength and other characteristics of an oscillating mode. Typical computational time for one variant amounts to several tens of minutes on IBM PC.

Acknowledgments

The authors appreciate fruitful discussions with Dr. V.I. Kozlovsky of Lebedev Physical Institute, Russia. Work is partially supported by Russian Foundation for Basic Research, project No. 08-02-00796-a.

References

1. Keller, J.B., Antman, S. (eds.): Bifurcation theory and nonlinear eigenvalue problems. W.A. Benjamin, Inc., New York (1969)
2. Rao, H., Steel, M.J., Scarmozzino, R., Osgood Jr., R.M.: High-power single-mode antiresonant reflecting optical waveguide-type vertical-cavity surface-emitting lasers. *IEEE J. Quantum Electron.* 37, 1435–1440 (2001)
3. Elkin, N.N., Napartovich, A.P., Vysotsky, D.V., Lavrushin, B.M., Kozlovsky, V.I.: Modeling of a Vertical Cavity Surface Emitting Laser with a Resonant Array of Quantum Wells. In: *AIP Conference Proc.*, vol. 1168, pp. 436–439 (2009)
4. Saad, Y.: *Numerical Methods for Large Eigenvalue Problem*. Manchester University Press, Manchester (1992)
5. Hadley, G.R.: Modeling of diode laser arrays. In: Botez, D., Scifres, D.R. (eds.) *Diode Laser Arrays*, ch. 4, pp. 1–72. Cambridge Univ. Press, Cambridge (1994)
6. Siegman, A.E.: Quasi fast Hankel transform. *Optics Letters* 1, 13–15 (1977)
7. Fox, A.G., Li, T.: Effect of gain saturation on the oscillating modes of optical masers. *IEEE Journal of Quantum Electronics QE-2*, 774–783 (1966)
8. Broyden, C.G.: A Class of Methods for Solving Nonlinear Simultaneous Equations. *Mathematics of Computation* 19(92), 577–593 (1965)
9. Vysotsky, D.V., Elkin, N.N., Napartovich, A.P., Kozlovsky, V.I., Lavrushin, B.M.: Simulation of a longitudinally electron-beam-pumped nanoheterostructure semiconductor laser. *Quantum Electronics* 39, 1028–1032 (2009)

Analysis of the CBS Constant for Quadratic Finite Elements

Ivan Georgiev¹, Maria Lymbery², and Svetozar Margenov²

¹ Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria
john@parallel.bas.bg

² Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 25A, 1113 Sofia, Bulgaria
mariq@parallel.bas.bg, margenov@parallel.bas.bg

Abstract. We study the behavior of the CBS constant as a quality measure for hierarchical two-level splittings of quadratic FEM stiffness matrices. The article is written in the spirit of [3] where the focus is on the robustness with respect to mesh and coefficient anisotropy. The considered splittings are: Differences and Aggregates (DA); First Reduce (FR); and hierarchical P-decomposition (P). The presented results show sufficient conditions for the existence of optimal order Algebraic Multi-Level Iteration (AMLI) preconditioners.

1 Introduction

Let us consider the elliptic boundary value problem

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x) \quad \text{in } \Omega, \quad (1)$$

$$u = 0 \quad \text{on } \Gamma_D, \quad (2)$$

$$(a(x)\nabla u(x)) \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N, \quad (3)$$

where Ω is a polygonal convex domain in R^2 and $f(x)$ is a given function in $L_2(\Omega)$. The coefficient matrix $a(x)$ is symmetric positive definite and uniformly bounded in Ω , $\mathbf{n}$ is the outward unit vector normal to the boundary $\Gamma = \partial\Omega$ and $\Gamma = \Gamma_D \cup \Gamma_N$. The related weak formulation reads as follows. For $f \in L_2(\Omega)$ find $u \in V \equiv H_D^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$ satisfying

$$\mathcal{A}(u, v) = (f, v) \quad \forall v \in H_D^1(\Omega), \quad \mathcal{A}(u, v) := \int_{\Omega} a(x)\nabla u(x) \cdot \nabla v(x) dx. \quad (4)$$

The domain Ω is assumed to be discretized by the partition $\mathcal{T}_h$ which is obtained by a proper refinement of a given coarser partition $\mathcal{T}_H$. Let $\mathcal{T}_H$ be aligned with the discontinuities of $a(x)$ so that over each element $e \in \mathcal{T}_H$ the functions $a_{i,j}(x)$ are smooth.

The variational problem (4) is discretized using the finite element method, i.e., the space V is replaced by a finite dimensional subspace V_h . Then the finite element formulation can be expressed by finding $u_h \in V_h$, satisfying

$$\mathcal{A}_h(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \quad \mathcal{A}_h(u_h, v_h) = \sum_{e \in \mathcal{T}_h} \int_e a(e) \nabla u_h \cdot \nabla v_h dx. \quad (5)$$

Here $a(e)$ is a piecewise constant symmetric positive definite matrix, defined by integral averaged values of $a(x)$ over each element from the coarser triangulation $\mathcal{T}_H$. The resulting discrete problem to be solved is a linear system of equations:

$$A_h u_h = F_h, \quad (6)$$

where A_h is the corresponding global stiffness matrix, F_h is the global right hand side and h is the mesh parameter for the underlying partition $\mathcal{T}_h$ of Ω .

2 Background Studies

Here we present some needed background to the problem (see, e.g. [1,3]). The analysis for an arbitrary triangle $e \in \mathcal{T}_h$ with coordinates $(x_i, y_i), i = 1, 2, 3$ can be performed on the reference triangle $\tilde{e}$ with coordinates $(0, 0), (1, 0), (0, 1)$. Transforming the finite element functions between these triangles, the bilinear form $\mathcal{A}_e(\cdot, \cdot)$ becomes:

$$\begin{aligned} \mathcal{A}_{\tilde{e}}(\tilde{u}, \tilde{v}) = \mathcal{A}_e(u(\tilde{x}, \tilde{y}), v(\tilde{x}, \tilde{y})) &= \int_{\tilde{e}} \left[\frac{\partial \tilde{u}}{\partial \tilde{x}}, \frac{\partial \tilde{u}}{\partial \tilde{y}} \right] \begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) \\ (x_3 - x_1) & (y_3 - y_1) \end{bmatrix}^{-1} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ &\times \begin{bmatrix} (x_2 - x_1) & (x_3 - x_1) \\ (y_2 - y_1) & (y_3 - y_1) \end{bmatrix}^{-1} \left[\frac{\partial \tilde{v}}{\partial \tilde{x}}, \frac{\partial \tilde{v}}{\partial \tilde{y}} \right]^T \left| \frac{\partial(x, y)}{\partial(\tilde{x}, \tilde{y})} \right| d\tilde{e} = \int_{\tilde{e}} \sum_{i,j} \tilde{a}_{ij} \frac{\partial \tilde{u}}{\partial \tilde{x}_i} \frac{\partial \tilde{v}}{\partial \tilde{x}_j} d\tilde{e}, \quad (7) \end{aligned}$$

where $\tilde{x} \geq 0$, $\tilde{y} \geq 0$, $\tilde{x} + \tilde{y} \leq 1$, and the coefficients $\tilde{a}_{ij}$ depend on both the angles of e and the coefficients a_{ij} of the diffusion matrix. Therefore, if a local analysis is applied, it suffices to consider the reference triangle and arbitrary anisotropic coefficient matrix $a(e)$, or alternatively, the isotropic Laplace operator and an arbitrary triangle e . In what follows we apply the second variant when studying estimates of the constant in the strengthened Cauchy-Bunyakowski-Schwarz inequality.

The global stiffness matrix A_h can be written in the form

$$A_h = \sum_{e \in \mathcal{T}_h} R_e^T A_e R_e, \quad (8)$$

where A_e is the element stiffness matrix and R_e is the restriction mapping of the global vector corresponding to the element $e \in \mathcal{T}_h$.

In this article we study the case of quadratic finite elements. The next theorem provides a simple geometric interpretation of the related element stiffness matrix.

Theorem 1. *The element stiffness matrix A_e in the case of quadratic finite elements corresponding to the Laplace operator and an arbitrary triangle e can be written in the form:*

$$A_e = \begin{bmatrix} \frac{b+c}{2} & -\frac{2c}{3} & \frac{c}{6} & 0 & \frac{b}{6} & -\frac{2b}{3} \\ -\frac{2c}{3} & \frac{4(a+b+c)}{3} & -\frac{2c}{3} & -\frac{4b}{3} & 0 & -\frac{4a}{3} \\ \frac{c}{6} & -\frac{2c}{3} & \frac{a+c}{2} & -\frac{2a}{3} & \frac{a}{6} & 0 \\ 0 & -\frac{4b}{3} & -\frac{2a}{3} & \frac{4(a+b+c)}{3} & -\frac{2a}{3} & -\frac{4c}{3} \\ \frac{b}{6} & 0 & \frac{a}{6} & -\frac{2a}{3} & \frac{a+b}{2} & -\frac{2b}{3} \\ -\frac{2b}{3} & -\frac{4a}{3} & 0 & -\frac{4c}{3} & -\frac{2b}{3} & \frac{4(a+b+c)}{3} \end{bmatrix},$$

where a , b and c equal the cotangents of the angles in $e \in \mathcal{T}_h$.

Proof. We consider the bilinear form

$$\mathcal{A}_e(u, v) = \int_e (u_x v_x + u_y v_y) de \quad (9)$$

for a given arbitrary non-degenerate triangle e . Without loss of generality we can assume that $\theta_1 = \max\{\theta_1, \theta_2, \theta_3\}$, where θ_i , $i = 1, 2, 3$ are the angles of the triangle as shown in Fig. 1. Let us introduce the notations $h = |OA|$, $p = |OB|$,

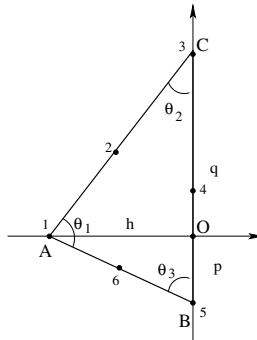


Fig. 1. Derivation of the element stiffness matrix

$q = |OC|$, $a = \cot\theta_1$, $b = \cot\theta_2$, $c = \cot\theta_3$. The next relations are readily seen:

$$b = \frac{p}{h}, \quad c = \frac{q}{h}, \quad a = \cot(\pi - (\theta_2 + \theta_3)) = \frac{h^2 - pq}{h(p + q)}. \quad (10)$$

Then the element stiffness matrix is derived by direct computation.

3 Hierarchical Two-Level Splittings

Let us consider a sequence of nested meshes $\mathcal{T}_H = \mathcal{T}_0 \subset \mathcal{T}_1 \subset \dots \subset \mathcal{T}_\ell = \mathcal{T}_h$. A uniform refinement procedure is used, i.e., the current coarse triangle $e \in \mathcal{T}_k$ is subdivided into four congruent triangles by joining the mid-edge nodes to obtain the macro-element $E \in \mathcal{T}_{k+1}$ as shown in Fig. 2. Let us denote by $A^{(0)}$, $A^{(1)}$, $\dots$, $A^{(\ell)}$ and by $\tilde{A}^{(0)}$, $\tilde{A}^{(1)}$, $\dots$, $\tilde{A}^{(\ell)}$ the related (standard basis) stiffness matrices and hierarchical basis stiffness matrices.

Preconditioners based on various multilevel extensions of two-level finite element methods lead to iterative methods which often have an optimal order of computational complexity with respect to the number of degrees of freedom of such a system. The key role in the derivation of optimal convergence rate estimates is played by the constant γ in the strengthened Cauchy-Bunyakowski-Schwarz (CBS) inequality, associated with the angle between the two subspaces of the splitting.

Here we focus our attention on the multiplicative AMLI preconditioner, $M_F = M_F^{(\ell)}$ defined recursively by:

$$M_F^{(0)} = A^{(0)}, \quad M_F^{(k)} = \begin{bmatrix} C_{11}^{(k)} & 0 \\ \tilde{A}_{21}^{(k)} & C_{22}^{(k)} \end{bmatrix} \begin{bmatrix} I & C_{11}^{(k)-1} \tilde{A}_{12}^{(k)} \\ 0 & I \end{bmatrix}, \quad (11)$$

where $C_{11}^{(k)}$ is some proper approximation the pivot block of the (hierarchical) stiffness matrix $\tilde{A}^{(k)}$ and the matrix $C_{22}^{(k)-1}$ is implicitly defined by the equation

$$C_{22}^{(k)-1} = [I - P_\mu(M_F^{(k-1)})^{-1} \tilde{A}^{(k-1)}] \tilde{A}^{(k-1)-1}, \quad (12)$$

where μ stands for the degree of the stabilization polynomial P_μ . In the case of regular refinement in 2D, the AMLI method has optimal computational complexity (for more details see [3]) if

$$(1 - \gamma^2)^{-1/2} < \mu < 4. \quad (13)$$

In what follows we introduce three hierarchical splittings for quadratic finite elements. Following (13), the robustness with respect to the anisotropy is studied based on the locally computed estimate of the CBS constant γ .

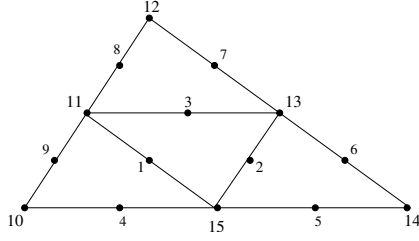


Fig. 2. Uniform refinement of a quadratic triangle element

3.1 Differences and Aggregates (DA) Splitting

Consider two consecutive meshes $\mathcal{T}_k \subset \mathcal{T}_{k+1}$. Let $\Phi_e^{(k)} = \{\phi_{e:i}^{(k)}\}_{i=1}^6$ and $\Phi_E^{(k+1)} = \{\phi_{E:i}^{(k+1)}\}_{i=1}^{15}$ be the standard finite element nodal basis functions for $e \in \mathcal{T}_k$ and $E \in \mathcal{T}_{k+1}$, see Fig. 2. We split the meshpoints $\mathcal{N}_E$ of E into two groups $\mathcal{N}_E = \mathcal{N}_e \cup \mathcal{N}_{E \setminus e}$, where $\mathcal{N}_e$ contains the common nodes for e and E . The Differences and Aggregates (DA) hierarchical basis is introduced as follows, see [2,3]:

$$\tilde{\Phi}_E^{(k+1)} = \{\phi_i^{(k)}\}_{i=1}^6 \cup \{\phi_j^{(k+1)}\}, \quad j \in \mathcal{N}_{E \setminus e}.$$

Then the local transformation matrix J_E , such that $\tilde{\Phi}_E^{(k+1)} = J_E \Phi_E^{(k+1)}$, has the form

$$J_E = \begin{bmatrix} I_9 \\ J_{E:21} & I_6 \end{bmatrix}, \quad J_{E:21} = \frac{1}{8} \begin{bmatrix} 0 & -1 & -1 & 3 & -1 & 0 & 0 & -1 & 3 \\ 4 & 2 & 4 & 0 & 0 & 0 & 0 & 6 & 6 \\ -1 & -1 & 0 & 0 & 0 & -1 & 3 & 3 & -1 \\ 2 & 4 & 4 & 0 & 0 & 6 & 6 & 0 & 0 \\ -1 & 0 & -1 & -1 & 3 & 3 & -1 & 0 & 0 \\ 4 & 4 & 2 & 6 & 6 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (14)$$

where I_9 and I_6 stand for the related identical matrix. The macro-element stiffness matrix A_E and the hierarchical basis matrix $\tilde{A}_E$ are related by $\tilde{A}_E = J_E A_E J_E^T$. Then the global hierarchical stiffness matrix $\tilde{A}^{(k+1)}$ can be assembled from the macro-element matrices $\tilde{A}_E^{(k+1)}$. Let us write the macroelement and the global matrices in the following 2×2 block form

$$A_E^{(k+1)} = \begin{bmatrix} A_{E:11}^{(k+1)} & A_{E:12}^{(k+1)} \\ A_{E:21}^{(k+1)} & A_{E:22}^{(k+1)} \end{bmatrix}, \quad \tilde{A}_E^{(k+1)} = \begin{bmatrix} \tilde{A}_{E:11}^{(k+1)} & \tilde{A}_{E:12}^{(k+1)} \\ \tilde{A}_{E:21}^{(k+1)} & \tilde{A}_{E:22}^{(k+1)} \end{bmatrix} \quad (15)$$

where the block $\tilde{A}_{E:22}^{(k+1)}$ is a 6×6 aggregated matrix corresponding to the nodal unknowns associated with the coarser mesh $\mathcal{T}_k$. The DA splitting is defined by (15) and is characterized by the related CBS constant γ_{DA} . One can prove that

$$A_{E:11}^{(k+1)} = \tilde{A}_{E:11}^{(k+1)}, \quad S_E^{(k+1)} = \tilde{S}_E^{(k+1)}, \quad \tilde{A}_{E:22}^{(k+1)} = A_e^{(k)}, \quad (16)$$

where $S_E^{(k+1)}$ and $\tilde{S}_E^{(k+1)}$ denote the local Schur complements for $A_E^{(k+1)}$ and $\tilde{A}_E^{(k+1)}$ respectively. Consequently $\ker(\tilde{A}_{E:22}^{(k+1)}) = \ker(A_e^{(k)}) = \text{span}\{(1, 1, 1, 1, 1)^T\}$, which enables us to apply a local analysis to the CBS constant γ_{DA} . From the general theory we get the relations

$$\gamma_{DA} \leq \max_{E \in \mathcal{T}_{k+1}} \gamma_{DA,E}, \quad \gamma_{DA,E}^2 = 1 - \mu_1, \quad (17)$$

μ_1 is the minimal eigenvalue of the generalized eigenproblem

$$\tilde{S}_E^{(k+1)} v_{E:2} = \mu A_e^{(k)} v_{E:2}, \quad v_{E:2} \neq (c, c, c, c, c)^T. \quad (18)$$

3.2 First Reduce (FR) Splitting

Similarly to DA, the FR splitting is introduced using the same macroelement transformation matrix J_E defined in (14). Then the hierarchical basis stiffness matrix $\tilde{A}^{(k+1)}$ is written in a (3×3) form

$$\tilde{A}^{(k+1)} = \begin{bmatrix} \tilde{A}_{11}^{(k+1)} & \tilde{A}_{12}^{(k+1)} & \tilde{A}_{13}^{(k+1)} \\ \tilde{A}_{21}^{(k+1)} & \tilde{A}_{22}^{(k+1)} & \tilde{A}_{23}^{(k+1)} \\ \tilde{A}_{31}^{(k+1)} & \tilde{A}_{32}^{(k+1)} & \tilde{A}_{33}^{(k+1)} \end{bmatrix}. \quad (19)$$

Here, the pivot block $\tilde{A}_{11}^{(k+1)}$ corresponds to the interior nodes of the macroelements $E \in \mathcal{T}_{k+1}$, the second diagonal block $\tilde{A}_{22}^{(k+1)}$ corresponds to the nodes from $\mathcal{N}_{E \setminus e}$ which are on the sides of the macroelements, and the last diagonal block is equal to the last diagonal block of the DA splitting and is therefore associated with the coarser mesh $\mathcal{T}_k$. Then the unknowns related to $\tilde{A}_{11}^{(k+1)}$ are first eliminated and the system with $\tilde{A}^{(k+1)}$ is reduced to a system with its Schur complement

$$\begin{aligned} B^{(k+1)} &= \begin{bmatrix} \tilde{A}_{22}^{(k+1)} & \tilde{A}_{23}^{(k+1)} \\ \tilde{A}_{32}^{(k+1)} & \tilde{A}_{33}^{(k+1)} \end{bmatrix} - \begin{bmatrix} \tilde{A}_{21}^{(k+1)} \\ \tilde{A}_{31}^{(k+1)} \end{bmatrix} [\tilde{A}_{11}^{(k+1)}]^{-1} \begin{bmatrix} \tilde{A}_{12}^{(k+1)} & \tilde{A}_{13}^{(k+1)} \end{bmatrix} = \\ &= \begin{bmatrix} B_{11}^{(k+1)} & B_{12}^{(k+1)} \\ B_{21}^{(k+1)} & B_{22}^{(k+1)} \end{bmatrix}. \end{aligned} \quad (20)$$

The FR splitting is defined by the 2×2 presentation of $B^{(k+1)}$. The block $B_{22}^{(k+1)}$ is associated with the coarse grid.

Let us note that $\tilde{A}_{11}^{(k+1)}$ is a block-diagonal matrix which allows to eliminate the interior unknowns locally. Therefore, we can assemble the Schur complement $B^{(k+1)}$ by the local ones $B_E^{(k+1)}$ where

$$B_E^{(k+1)} = \begin{bmatrix} B_{E:11}^{(k+1)} & B_{E:12}^{(k+1)} \\ B_{E:21}^{(k+1)} & B_{E:22}^{(k+1)} \end{bmatrix}. \quad (21)$$

One can prove again that $\ker(B_{E:22}^{(k+1)}) = \ker(A_e) = \text{span}(1, 1, 1, 1, 1)^T$. Similarly to (17) and (18) we can estimate the CBS constant γ_{FR} using the locally computed $\gamma_{FR,E}$, corresponding to the splitting (21).

3.3 Hierarchical Basis P Splitting

In this subsection we briefly present the hierarchical two-level splitting (first analyzed in [4]) which makes use of both piecewise linear and piecewise quadratic basis functions. Let us consider the linear FEM discretization corresponding to the triangulation $\mathcal{T}_k$. Then at the refinement step we keep the piecewise linear basis functions at the vertex nodes adding piecewise quadratic functions to the mid-edge nodes.

According to this P hierarchical splitting of the unknowns we present the macro-element stiffness matrix $\bar{A}_E^{(k+1)}$ and the assembled global stiffness matrix $\bar{A}^{(k+1)}$ in the block form

$$\bar{A}_E^{(k+1)} = \begin{pmatrix} \bar{A}_{E:11}^{(k+1)} & \bar{A}_{E:12}^{(k+1)} \\ \bar{A}_{E:21}^{(k+1)} & \bar{A}_{E:22}^{(k+1)} \end{pmatrix}, \quad \bar{A}^{(k+1)} = \begin{pmatrix} \bar{A}_{11}^{(k+1)} & \bar{A}_{12}^{(k+1)} \\ \bar{A}_{21}^{(k+1)} & \bar{A}_{22}^{(k+1)} \end{pmatrix}. \quad (22)$$

The second diagonal blocks of both $\bar{A}_E^{(k+1)}$ and $\bar{A}^{(k+1)}$, correspond to the linear finite elements defined on the coarser mesh $\mathcal{T}_k$. Therefore, $\ker(\bar{A}_{E:22}^{(k+1)}) = \text{span}(1, 1, 1)^T$, and local analysis can be applied to compute the macroelement CBS constant $\gamma_{E:P}$, and thereby to estimate the global constant γ_P .

4 Numerical Study of the CBS Constants

The construction of robust two-level methods for higher order FEM problems with respect to mesh and/or coefficient anisotropy is still an open problem. In this section we present a comparative numerical study of the DA, FR, and P hierarchical splittings. In the presented local analysis, without loss of generality, we can assume that the angles θ_1 , θ_2 and θ_3 of the arbitrary element e satisfy the condition $\theta_1 \geq \theta_2 \geq \theta_3$. Therefore if a , b and c equal the cotangents of the angles, we have that, see e.g. [3], $|a| \leq b \leq c$, $a = (1 - bc)/(b + c)$. Then by setting $\alpha = a/c$ and $\beta = b/c$ we can estimate the local CBS constants in terms of (α, β) whose admissible domain D is given by

$$D = \left\{ (\alpha, \beta) \in \mathcal{R}^2 : -\frac{1}{2} < \alpha \leq 1, \max\left\{-\frac{\alpha}{\alpha+1}, |\alpha|\right\} \leq \beta \leq 1 \right\}. \quad (23)$$

The sets of $\{\alpha, \beta\}$ for which the local CBS constants satisfy the inequality $\gamma_E^2 < \frac{3}{4}$ are shown in Fig. 3. According to the AMLI optimality condition (13), this case corresponds to a stabilization polynomial of degree $\mu = 2$. Similarly Fig. 4 shows the domain of the same parameters for which we have $\gamma_E^2 < \frac{8}{9}$, i.e. $\mu = 3$.

We can observe that the region subtended by the FR splitting is always bigger than the region subtended by the DA splitting. This is in full agreement with the theory of generalized FR splittings, i.e., $\gamma_{FR} \leq \gamma_{DA}$. We see also the general advantage of FR for problems with stronger anisotropy (see also Table 1). However, for some cases of more modest anisotropy, we get $\gamma_P \leq \gamma_{FR}$. (see, e.g., [3]).

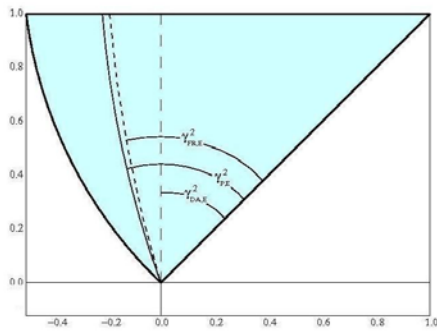


Fig. 3. $\{\alpha, \beta\} : \gamma_E^2 \leq \frac{3}{4}$

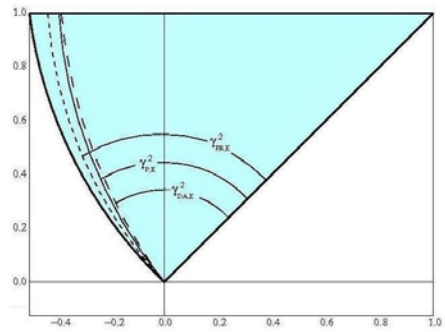


Fig. 4. $\{\alpha, \beta\} : \gamma_E^2 \leq \frac{8}{9}$

Table 1. γ_E^2 for isosceles triangles

θ_1	θ_2	θ_3	$\gamma_{DA,E}^2$	$\gamma_{FR,E}^2$	$\gamma_{P,E}^2$
100°	40°	40°	0.7913	0.7265	0.7245
120°	30°	30°	0.8598	0.8024	0.8333
140°	20°	20°	0.9086	0.8836	0.9220
160°	10°	10°	0.9490	0.9490	0.9798

On the basis of the obtained computational results by the software package Mathematica, we conclude that for a fixed minimal angle θ_3 , the largest CBS constant corresponds to the case of an isosceles triangle with $\theta_2 = \theta_3$. This is a motivation for the selection of data presented in Table 1. We see in particular, that for the FR splitting the AMLI method with $\mu = 3$ satisfies the optimality condition (13) if the minimal angle $\theta_3 \geq 20^\circ$. Let us note that such kind of conditions can be controlled by many of the available advanced mesh generators.

Acknowledgement. The partial support of the Bulgarian NSF Grants DO 02-115/08 and DO 02-338/08 is highly appreciated.

References

1. Axelsson, O.: Stabilization of Algebraic Multilevel Iteration method; additive methods. Numerical Algorithms, 23–47 (1999)

2. Blaheta, R., Margenov, S., Neytcheva, M.: Uniform Estimate of the Constant in the Strengthened CBS inequality for Anisotropic Non-conforming FEM systems. Numerical Linear Algebra and Applications 11(4), 309–326 (2004)

3. Kraus, J., Margenov, S.: Robust Algebraic Multilevel Methods and Algorithms. De Gruyter, Germany (2009)

4. Maitre, J.F., Musy, S.: The Contraction Number of a Class of Two-level Methods; An Exact Evaluation for Some Finite Element Subspaces and Model Problems. Lect. Notes Math., vol. 960, pp. 535–544 (1982)

Sensitivity of Results of the Water Flow Problem in a Discrete Fracture Network with Large Coefficient Differences

Milan Hokr, Jiří Kopal, Jan Březina, and Petr Rálek

Technical University of Liberec, Studentská 2, Liberec, 46117, Czech Republic
`milan.hokr@tul.cz`

Abstract. This work deals with modelling of groundwater flow in compact rock with network of discrete fractures. The test problem is given by stochastically generated network of lines (fractures) with large variations of the aperture, conductivity, and discretisation element size which leads to the differences of the coefficients in the linear equations system up to ten orders of magnitude. We compare our own simulation code using mixed finite elements with commercial code NAPSAC using standard finite elements. Both codes produce consistent results, with differences in percents but unevenly distributed. Results from mixed finite elements have four orders of magnitude smaller error of mass balance than those from standard finite elements.

1 Introduction

Modelling of groundwater flow and other physical processes in rock material is among the well-known field of application of numerical methods as well as a source of the particular problems motivating further research and improvements in the numerical mathematics. The modelling tasks come from various industrial and environmental problems; the topic presented in this paper is related to safety analysis of the deep geological repository of spent nuclear fuel, where the need to precisely predict the radionuclide migration is generally declared.

Typically the discretisation methods for partial differential equations are known to worsen the stability for large spatial differences of the coefficients in the equation (inhomogeneous material) and/or for large differences of discretisation parameter (element size). Even if this is cited in most of the textbooks and basic courses, not so frequently it is seen in practical problems; especially this happens in the context of adaptivity and for large problems (measured by the degrees of freedom). In this paper we present particular demonstration of this phenomenon, together with comparison how two different variants of finite element method can behave differently in such conditions. The solved problem has fixed discretisation geometry, where the coefficient inhomogeneity comes from physical nature of the solved problem (inhomogeneity of rock material) and the discretisation differences come from stochastic origin of the problem geometry (intersections of fractures with position generated stochastically).

The reason why it is studied and its scientific role in the rock flow and transport modelling is that it represents more realistically the inhomogeneity of flow velocity and solute particles distribution – there are several main “channels” dominating with its velocity and total flux and a lot of smaller fractures contributing to the retention capacity of the rock (visible in Fig. 5). Correct estimation of the distribution of transport velocity is thus important to assess the possible retardation of the radionuclides in our case, by e.g. sorption or matrix diffusion.

2 Problem Description

The fracture network was generated from the site geological mapping of Sellafield, UK [6]. For the case of numerical simulation benchmarking, a planar problem of 1D fractures set (representing a cross-section of real 3D rock with planar fractures) is defined: as the square 10×10 m with 7786 fractures. The fractures are defined by their position (ending point coordinates), length, and aperture (meaning the thickness in geological terminology). According to geological observation, the length and aperture are correlated, i.e. larger fractures are larger by both their length and aperture and vice versa.

Here, the fracture thickness is understood in the direction of the 2D model, i.e. the fracture is a 2D strip in the 2D plane with its physical meaning, but it is represented by a 1D line in the model geometry (Fig. 1). With respect to the 3D world, the 2D model is considered with unit thickness.

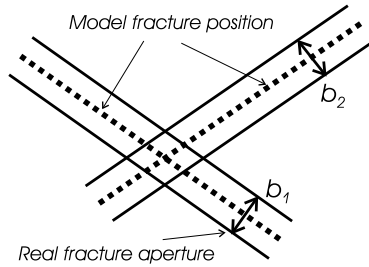


Fig. 1. Illustration how the fracture thickness (aperture) is understood with respect to its representation by a line in the model

The solved phenomenon, water flow in a narrow channel, is described for a single fracture by the potential flow model,

$$\mathbf{u} = -K \nabla p \quad (1)$$

$$\nabla \cdot \mathbf{u} = q, \quad (2)$$

where $\mathbf{u}(x, t)$ [m/s] is the unknown velocity, $p(x, t)$ [m] is unknown pressure head (pressure represented in metres of water column, i.e. $p = \frac{\tilde{p}}{\rho g}$, where $\tilde{p}$ [Pa])

is pressure, ϱ is the density, and g is the gravity acceleration), K is the hydraulic conductivity, and q are sources/sinks (zero in our model). In places of fracture intersections, we assume continuity of pressure head and mass balance of fluxes.

The hydraulic conductivity is governed by $K = \frac{\varrho g}{12\mu} b^2$ (called Hagen-Poiseuille or “cubic” law, as the flux $b \mathbf{u}$ [m²/s] is proportional to b^3), where μ is the dynamic viscosity [Pa s]. Thus the aperture b is the only input parameter besides the fracture network geometry and boundary conditions.

The Dirichlet boundary conditions are defined over whole boundary, which means prescribed pressure head in the end points of the fractures lying on the model square boundary. Zero Neumann condition (no flow) is considered at the end points of the fractures inside the model (the problem can be equivalently solved without the dead-end parts of fractures, i.e. when no such end points exist). The values of the Dirichlet boundary condition represent a uniform pressure gradient (of unit magnitude) in either horizontal (x) or vertical (y) directions, i.e. pressure is constant on perpendicular boundaries and a linear function on parallel boundaries (Fig. 2):

- for horizontal gradient $p(x, y) = 20 + x$ (ranging between $p_1 = 10$ and $p_2 = 30$)
- for vertical gradient $p(x, y) = 20 + y$ (ranging between $p_1 = 10$ and $p_2 = 30$)

For further text we will reference the boundaries perpendicular to gradient, i.e. those with constant boundary pressure, as inflow and outflow boundary.

We use five different distributions of aperture on the same geometry of the fracture network. Actually they are evaluated from the reference case as the influence of variable ratio of horizontal and vertical mechanical stress [4,3], but it is not important for the study presented in this paper (we will use the stress ratio almost like a “meaningless” notation of different fracture aperture distribution variants).

Table 1. Geometrical and material parameters associated with discretisation elements, for the reference case (no mechanical stress)

	K	b	Δx	$\frac{Kb}{\Delta x}$ real	$\frac{Kb}{\Delta x}$ worst possible
Maximum	2.4e-2	2e-4	8.4e-1	2.6e-1	$\frac{\max Kb}{\min \Delta x} = 2.4$
Minimum	1.1e-5	4.1e-6	2e-6	1.1e-10	$\frac{\min Kb}{\max \Delta x} = 5.4e-11$
Max./Min.	2.2e+3	4.7e+1	4.1e+5	2.4e+9	4.5e+10

3 Numerical Solution

We use two different variants of finite element method implemented in two simulation codes of different kind. The first one is the code FLOW123D developed by the authors’ team at the Technical University of Liberec. The numerical method used is the mixed-hybrid finite element method and the main feature is the

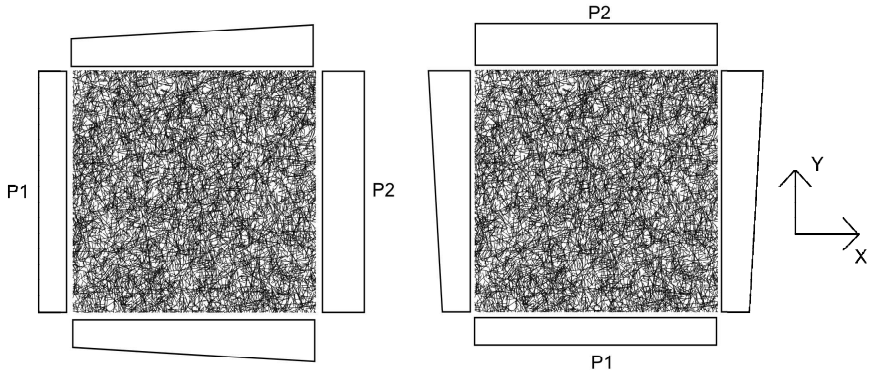


Fig. 2. Distribution of the Dirichlet boundary condition values (prescribed pressure head p) along the displayed fracture network used for calculation

multidimensional model geometry [5] – combination of subdomains of different geometric dimension (1D, 2D, 3D), representing either the fractures or the rock continuum in a single model. In each dimension the finite elements are formulated and the resulting system of equations is completed by the discrete form of the physical interaction between the different domains (e.g. between rock matrix and fracture). The Raviart-Thomas piece-wise linear base functions are used for the velocity approximation and piece-wise constant base function are used for the pressure approximation (together with the Lagrange multipliers representing the pressure on the element sides). The main feature of the method is direct calculation of velocity/flux and mass balance. The current developments of the algebraic solver are presented in this proceedings [1].

The second one, commercial code NAPSAC [2], uses the standard linear finite elements and it is one of the typical tools used by hydrogeologists for fractured rock problems. The simulations have been done by contract from a consulting company [7]. In the comparison, NAPSAC has a role of “established standard” providing a verification of our code FLOW123D, but it is also an example of code with some limitations resulting from the used numerical method (the mass balance error presented here).

The discretisation is only given by the fracture intersections, i.e. a segment of fracture (line) between two neighbour intersections is a discretisation element. There are in total 74826 such segments. The number reduces to 60052 segments if we delete the dead-end segments which do not contribute to flux. The size of the system of linear algebraic equations from the mixed-hybrid finite-element method (FLOW123D code) is 273570, with 773994 non-zeros in the system matrix.

Independently of the used numerical method, the values in the discretised problem, i.e. the system of linear algebraic equations, depend on the ratio $\frac{Kb}{\Delta x} \sim \frac{b^3}{\Delta x}$ (where Δx is the element size, i.e. the length of the segment between intersections). The ratios evaluated from the used fracture network are presented in

Tab. 1. The real maximum and minimum of the ratio (over all 60052 fracture segments) can be compared with ratios for worst possible combination (maxima and minima) of single parameters (aperture and length).

The problem conditioning could be improved by discretisation of the segments (with respect to the continuous physical problem we do not improve accuracy as the solution is a linear function of position), but the improvement of the ratio $\frac{Kb}{\Delta x} \sim \frac{b^3}{\Delta x}$ of the discrete problem is maximum one order of magnitude (if we keep the problem size withing reasonable limit) and we expect it is not worth of the rise of the degrees of freedom.

Table 2. Comparison of codes by total flux through Left, Right, Bottom, and Top boundary, for the horizontal pressure gradient

	Stress ratio				
	0 (none)	1	2	3	5
NAPSAC L	9.21E-05	2.41E-05	1.80E-05	1.46E-05	1.17E-05
NAPSAC R	-1.00E-04	-2.63E-05	-2.00E-05	-1.62E-05	-1.25E-05
NAPSAC B	2.73E-06	5.80E-07	1.37E-06	1.47E-06	1.36E-06
NAPSAC T	5.53E-06	1.72E-06	7.03E-07	1.61E-07	-4.49E-07
FLOW123D L	9.39E-05	2.46E-05	1.85E-05	1.51E-05	1.22E-05
FLOW123D R	-9.86E-05	-2.59E-05	-1.97E-05	-1.59E-05	-1.24E-05
FLOW123D B	-6.21E-07	-3.23E-07	5.40E-07	7.08E-07	6.32E-07
FLOW123D T	5.33E-06	1.69E-06	6.74E-07	1.34E-07	-4.74E-07
balance NAPSAC	2.07E-07	9.60E-08	1.04E-07	9.77E-08	1.34E-07
balance FLOW123D	-2.00E-11	1.30E-11	4.20E-11	1.80E-11	4.80E-11
rel. error L	-2.02E-02	-1.90E-02	-2.65E-02	-3.26E-02	-3.96E-02
rel. error R	1.37E-02	1.42E-02	1.59E-02	1.58E-02	1.12E-02
rel. error B	1.23E+00	1.56E+00	6.06E-01	5.18E-01	5.35E-01
rel. error T	3.58E-02	1.62E-02	4.17E-02	1.69E-01	-5.66E-02

4 Results and Comparison

We compare the results by means of total fluxes through each side of the model square and the distribution of flux along the side discretised to 100 segments of the length 0.2 m. The sign convention is positive for outflow and negative for inflow. Next the total water balance is evaluated, i.e. the sum of fluxes through all the four model edges, which should be ideally zero (the total inflow equal to the total outflow). The fluxes are compared in terms of relative difference $(q_{NAP} - q_{F123})/q_{NAP}$, where q is the flux through a particular model side with subscript denoting the code FLOW123D or NAPSAC. The data are presented in Tab. 2 for only the horizontal gradient due to the limited space.

The fit of both models/codes is generally good, with an exception mentioned below. The dominant flux is through the edges perpendicular to the pressure gradient (the right for inflow and the left for outflow in case of the horizontal gradient), while the fluxes through the lateral sides are one or two orders of

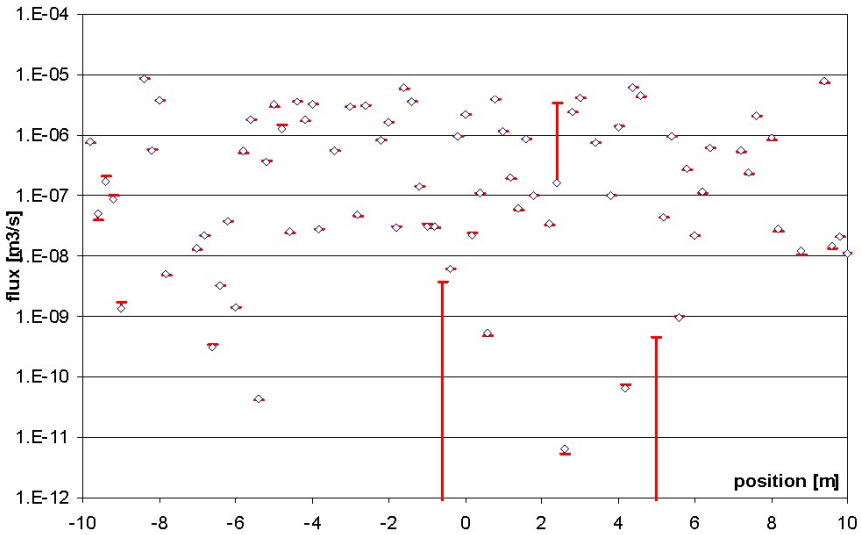


Fig. 3. Distribution of flux along the outflow (left for the horizontal gradient) boundary – the points represent the NAPSAC results and the error bars represent the differences to the FLOW123D results

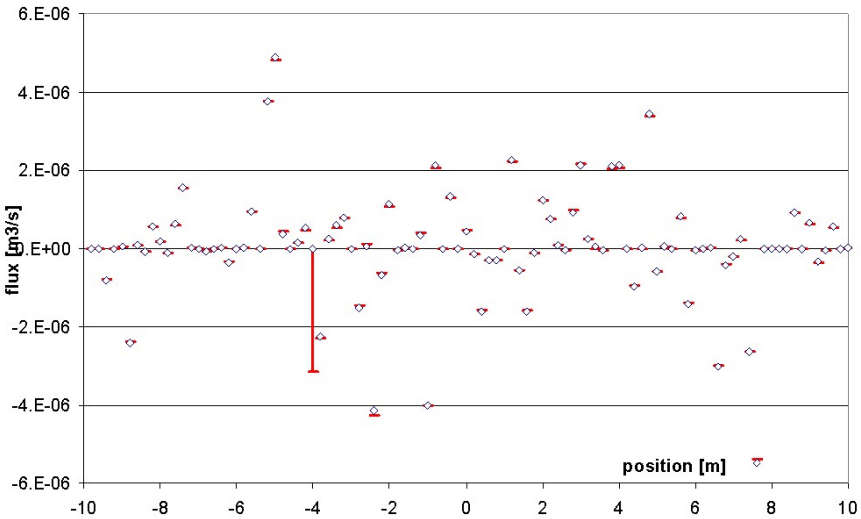


Fig. 4. Distribution of flux along the lateral (bottom for the horizontal gradient) boundary – the points represent the NAPSAC results and the error bars represent the differences to the FLOW123D results

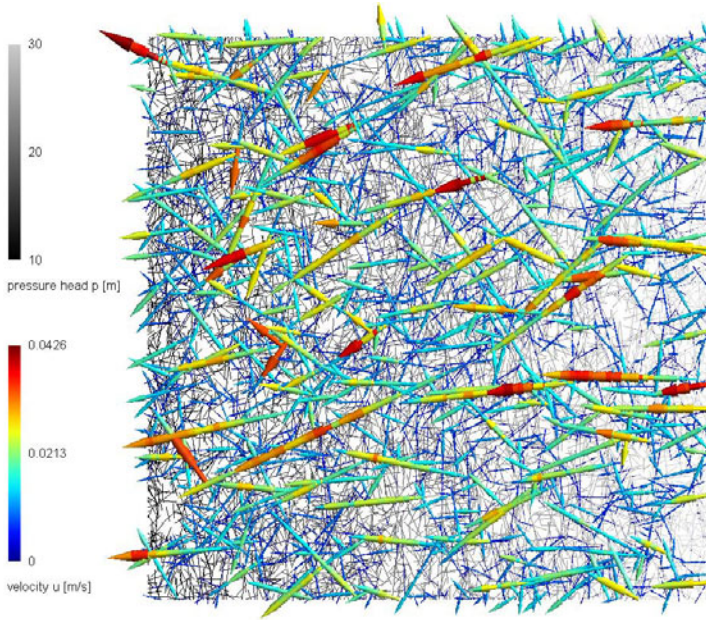


Fig. 5. Example model results – the pressure head changes almost uniformly in the left–right direction (gray levels on the background fracture drawing), the velocity is visible only in few of the fractures (arrows in the foreground, color in the online version)

magnitude smaller, governed by the fracture network anisotropy. The relative difference of models is in the order of percents for the inflow and the outflow boundary and for one of the lateral boundaries, but larger for the second lateral boundary (with the smallest absolute flux).

The fulfillment of the mass balance condition is distinctively better for the FLOW123D code, between 3 and 4 orders of magnitude. The balance error of the NAPSAC results is in the order of percents of the larger fluxes and sometimes comparable to the smallest fluxes. From another point of view, the balance error of NAPSAC is comparable with the difference between FLOW123D and NAPSAC in some cases (the side with smallest flux and largest difference for the vertical gradient which is not presented). All these observations can be understood as an argument for the FLOW123D results as more credible.

In total there are 40 profiles of flux distribution along the model side to analyse (4 sides, 2 variants of boundary gradient, and 5 stress states). In this paper of limited extend we show two profiles representing the typical results, for horizontal gradient, no-stress state, and one outflow boundary (larger total flux, all local fluxes of the same sign/direction) and one lateral boundary (smaller total flux, local fluxes of different sign/direction).

The comparison for the left (outflow) boundary is in Fig. 3. Graphically, the data fit well (the relative error is about a percent, similar as for the total flux Tab. 2), except three cases: one of them correspond to the displayed different

values, the two remaining represent a small non-zero value from FLOW123D versus zero value (not displayed on the logarithmic scale) for NAPSAC. The logarithmic scale is used to cover both the smaller and larger fluxes (most of them in four orders of magnitude range). The largest visible difference of $3.18\text{E-}06 \text{ m}^3/\text{s}$ dominates the error of the total flux through the side which is $1.87\text{E-}06 \text{ m}^3/\text{s}$ (compensated by more fractures of negative flux difference).

The comparison for the bottom (parallel to flow) boundary is in Fig. 4. There is one value with significant difference (non-zero versus zero) which fully dominates the total difference. The remaining values differ within the range of percents and also the positive and negative fluxes cancel in the sum.

5 Conclusion

The comparison gives good argument for FLOW123D verification and clearly demonstrates the mass-balance properties of mixed-hybrid finite element methods which is not destroyed by the algebraic round-off errors. Some part of the difference between NAPSAC and FLOW123D can be related to the mass balance error of NAPSAC which is in the same order of magnitude.

Acknowledgement. This research was supported by the Grant Agency of Czech Republic under project no. 205/09/1879.

References

1. Březina, J., Rálek, P., Hokr, M.: Parallel Simulator of Multidimensional Fracture Flow and Transport. In: NM&A (2010)
2. Hartley, L.J.: NAPSAC Release 4.1 Technical Summary Document, AEA-R&R-0271. AEA Technology (1998)
3. Havlíček, J., Hokr, M.: Change of the hydraulic parameters in the model of flow in discrete fracture network caused by the mechanical stress. In: Plešinger, M. (ed.) Simulation, Modelling, and Various Applications (SIMONA 2009), pp. 37–43. Technical University of Liberec (2009)
4. Jing, L., Hudson, J.A. (eds.): Task C: Integrated assessment of THMC coupled processes in single fractures and fractured rocks, DECOVALEX-2011 Project Progress Report, Stage1 (in preparation).
5. Maryška, J., Severýn, O., Tauchman, M., Tondr, D.: Modelling of Processes in Fractured Rock Using FEM/FVM on Multidimensional Domains. *J. Comp. Appl. Math.* 215(2), 495–502 (2008)
6. Min, K.B., Jing, L., Stephansson, O.: Determining the Equivalent Permeability Tensor for Fractured Rock Masses Using a Stochastic REV Approach: Method and Application to the Field Data from Sellafield, UK. *Hydrogeol. J.* 12(5), 497–510 (2004)
7. Polák, M., Milický, M., Gvoždík, L., Uhlík, J.: Flow simulation in 2D fracture network, Technical report, PROGEO Ltd. (2009) (in Czech)

Fluxon Dynamics in Stacked Josephson Junctions

Ivan Hristov and Stefka Dimova

Faculty of Mathematics and Informatics, University of Sofia,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
`ivanh@fmi.uni-sofia.bg`, `dimova@fmi.uni-sofia.bg`

Abstract. Sakai-Bodin-Pedersen model – a system of perturbed sine-Gordon equations – is used to study numerically the dynamics of Josephson phases in stacks of inductively coupled long Josephson Junctions (LJJs). The boundary conditions correspond to a stack of linear geometry. In order to obtain appropriate initial values for the dynamic problem the corresponding static problem is solved as well. We are interested in solutions having one or two moving fluxons in each junction and seek for conditions under which a bunching of fluxons is possible. The current-voltage dependencies and the current-velocity dependencies for different values of dissipation and coupling parameters for bunched and unbunched states are found. To solve numerically the above problems Finite element method and Finite difference method are used.

1 Introduction

In recent years, much attention has been attracted to different kinds of solid-state multilayered systems, for example Josephson and magnetic multilayers, high-temperature superconductors and perovskites. Multilayers are attractive because it is often possible to multiply a physical effect achieved in one layer by N , N being the number of layers. In addition, some new important physical effects as current locking and Cherenkov radiation by Josephson fluxons may occur because of the interaction between subsystems. The possibility of comparing theoretical predictions with experimental measurements increases the interest to these systems.

The fluxon dynamics and the possibility of fluxon bunching in stacked LJJs have been investigated during the last 17 years. Stacks of annular (ring) geometry or of infinite size are considered so as to avoid ambiguities due to reflection of the fluxons from the edges in the linear (open ends) geometry case. Fluxon bunching in single annular Josephson Junctions is investigated experimentally and numerically in [11]. The stability of bunched states in both inductively and capacitively coupled systems of two infinite size JJs is shown by numerical simulations [6] and compared with the predictions made on the basis of the fundamental bunched soliton solution to the corresponding unperturbed Sine-Gordon system. The motion of a fluxon in one or two magnetically coupled annular JJs is investigated experimentally and theoretically in [3]. The case of two and three inductively coupled annular junctions is analyzed in [2] for some different combinations of fluxons in

the system. A simple analytical expression, which shows the possibility bunched states to exist, is derived. The propagation of one fluxon in each junction of a system of three stacked annular JJs is analyzed in [4] and [5].

In this work we show by numerical experiment that bunching of fluxons in the case of three geometrically symmetric LJJs of linear geometry occurs. We investigate the solutions with one and two moving fluxon in each junction with respect to the coupling and dissipation parameters.

In the next section the mathematical model is described. In section 3 the numerical methods and algorithms are briefly discussed. The numerical results are described and shown in section 4. The last section contains the conclusions.

2 Mathematical Model

We use the inductive coupling model of Sakai - Bodin - Pedersen from [10], where a theory, describing the interaction between a general system of N junctions is deduced from the Maxwell, London and Josephson equations.

N -stacked LJJ consists of $N+1$ superconducting layers of thickness d , divided by N insulating layers of thickness D . In the symmetric case the electromagnetic interaction between junctions is represented by a coupling parameter S , $(-0.5 < S \leq 0)$, given by

$$S = -\lambda / (D \sinh(d/\lambda) + 2\lambda \cosh(d/\lambda)),$$

where λ is the London penetration depth [8]. A junction can be treated as one-dimensional, if its length is much bigger than the Josephson penetration length λ_J [8] and its width is smaller than λ_J .

In the case of symmetric three stacked inductively coupled long Josephson junctions, considered here, the dynamics of the Josephson phases $\varphi(x, t) = (\varphi_1(x, t), \varphi_2(x, t), \varphi_3(x, t))^T$ is described by the following system of perturbed sine-Gordon equations [10]:

$$\varphi_{tt} + \alpha \varphi_t + J + \Gamma = L^{-1} \varphi_{xx}, \quad -\ell < x < \ell, \quad 0 < t \leq T. \quad (1)$$

Here 2ℓ is the length of the stack, α is the dissipation coefficient (damping parameter), $\Gamma = \gamma(1, 1, 1)^T$ is the vector of the external current density, $J = (\sin \varphi_1, \sin \varphi_2, \sin \varphi_3)^T$ is the vector of the Josephson current density. The matrix $L = \text{tridiag}(1, S, 1)$ is the inductive coupling matrix, In system (1) the space x is normalized with respect to λ_J and the time t - to the inverse of the plasma frequency.

In this work we consider stacks of linear geometry placed in external magnetic field h_e , therefore the system (1) should be solved together with the boundary conditions:

$$\varphi_x(-\ell) = \varphi_x(\ell) = H, \quad (2)$$

where H is the vector $H = h_e(1, 1, 1)^T$. To close the differential problem appropriate initial conditions must be posed:

$$\varphi(x, 0) - \text{given}, \quad \varphi_t(x, 0) - \text{given}. \quad (3)$$

The most important solution of a single unperturbed one dimensional sine-Gordon equation

$$\varphi_{tt} - \varphi_{xx} + \sin \varphi = 0, \quad -\infty < x < \infty \quad (4)$$

is given by:

$$\varphi(x, t) = 4 \arctan\left[\exp\left(\sigma \frac{x - ut - x_0}{\sqrt{1 - u^2}}\right)\right],$$

where $\sigma = +1$ corresponds to a moving fluxon with velocity u , $\sigma = -1$ corresponds to a moving antifluxon and x_0 is the location of the fluxon at $t = 0$.

The important solutions of the corresponding to (4) static equation are:

- Meissner solutions, denoted by M : $\varphi(x) = k\pi$, $k = 0, \pm 1, \pm 2, \dots$,
- one-fluxon (antifluxon) solutions: $\varphi(x) = 4 \arctan(\exp(\pm x)) + 2k\pi$. For n -fluxon distributions, in both static and dynamic cases, the notation F^n is used.

We use the mentioned above static solutions to form initial conditions for the dynamic problem (1), (2), (3). In the three stacked case we consider static solutions, which are combinations of solutions existing in the one junction case. For example, the notation (F^1, F^1, F^1) is used for a distribution corresponding to one-fluxon in each junction. To find moving fluxons solutions we excite appropriate static solutions by increasing the external current γ .

The existence of Josephson current generates a specific magnetic flux. When the external current γ is less than some critical value, all the junctions are in some static state, i.e., we have a time independent solution of the system (1), (2). In this case the measured voltages in all junctions are zero. When this critical value is exceeded, the system switches to dynamic state and the voltage of at least one of the junctions becomes nonzero. The voltage in i -th junction is:

$$V_i = \lim_{T \rightarrow \infty} \frac{1}{2\ell T} \int_0^T \int_{-\ell}^{\ell} \varphi_{i,t}(t, x) dx dt. \quad (5)$$

We will need further the so called Swihart velocities [4], which for the three stacked JJs are

$$c_{\pm} = 1/\sqrt{1 \pm \sqrt{2}S}, \quad c_d = 1.$$

They appear in the analysis of the bunched fluxons states for stacked JJs in the case of annular geometry [4], [5]. We will compare with them the velocities of the bunched fluxons states in the case of linear geometry.

3 Numerical Methods and Algorithms

To solve the dynamic problem (1), (2), (3), we use the finite difference method. The main equation (1) is approximated by the "cross-shaped" scheme. To approximate the boundary conditions (2), two different approximations - of second order (Scheme 1) and third order (Scheme 2) - are used. Let h and τ be the steps in space and time respectively, $\delta = (\tau/h)^2$, n - the number of points in

space, $x_k = -\ell + kh$, $h = 2\ell/n$, $k = 0, \dots, n$, $t_j = j\tau$, $j = 0, 1, \dots$. By using the standard notations:

$$y_k^l = \varphi_l(x_k, t_j), \quad \tilde{y}_k^l = \varphi_l(x_k, t_{j+1}), \quad \check{y}_k^l = \varphi_l(x_k, t_{j-1}), \quad l = 1, \dots, N,$$

the difference scheme for three stacked LJJs ($N = 3$) can be written as follows:

$$\hat{y}_k^l = \frac{1}{1 + 0.5\alpha\tau} \left[2y_k^l + (0.5\alpha\tau - 1)\tilde{y}_k^l - \tau^2(\sin y_k^l + \gamma) + \sum_{m=1}^3 \delta a_{l,m} y_{\bar{x},k}^m \right], \quad (6)$$

where $l = 1, 2, 3$, $k = 1, \dots, n-1$, $L^{-1} = (a_{l,m})_{l,m=1}^3$; for Scheme 1:

$$\hat{y}_0^l = (4\hat{y}_1^l - \hat{y}_2^l - 2hh_e)/3, \quad \hat{y}_n^l = (4\hat{y}_{n-1}^l - \hat{y}_{n-2}^l + 2hh_e)/3; \quad (7)$$

for Scheme 2:

$$\hat{y}_0^l = (18\hat{y}_1^l - 9\hat{y}_2^l + 2\hat{y}_3^l - 6hh_e)/11, \quad \hat{y}_n^l = (18\hat{y}_{n-1}^l - 9\hat{y}_{n-2}^l + 2\hat{y}_{n-3}^l + 6hh_e)/11. \quad (8)$$

To check the numerical stability and the real order of accuracy of the schemes (6), (7) and (6), (8) we have made computations for fixed time levels and embedded meshes in space. The results show second order of convergence in space and time. In addition we verify the integral identity [7]

$$Q(t) = \frac{d}{dt}E(t) + \alpha \int_{-\ell}^{\ell} [2\varphi_{1,t}^2 + \varphi_{2,t}^2] dx = 0, \quad (9)$$

$$E(t) = \int_{-\ell}^{\ell} \left[\frac{\varphi_{1,x}^2 + \frac{1}{2}\varphi_{2,x}^2}{1 - 2S^2} + 3 - 2\cos\varphi_1 - \cos\varphi_2 - \gamma(2\varphi_1 + \varphi_2) \right] dx +$$

$$\int_{-\ell}^{\ell} \varphi_{1,t}^2 dx + \frac{1}{2}\varphi_{2,t}^2 dx + \frac{-2S}{1 - 2S^2} \int_{-\ell}^{\ell} \varphi_{1,x}\varphi_{2,x} dx.$$

The approximate value $\tilde{Q}(t)$ of (9) for the explicit in time schemes corresponding to the two types of approximations of the boundary conditions (7), (8) are shown on Fig. 1. For the two schemes the biggest values of $\tilde{Q}(t)$ are at a small time interval, when the fluxons reflect from the ends of the junctions and change their polarity. The fluxon reflection from edges results in a large energy dissipation due to a plasma wave emission [2], so this discrepancy is natural. Let us mention, outside this interval the values of $\tilde{Q}(t)$ are of order $10^{-7} - 10^{-8}$.

To find the approximate value of the voltage V_i given by (5), an averaging procedure is proposed and realized in [1]. Let us note that the calculation of the current-voltage characteristics is a much time consuming task. An algorithm for calculating the average fluxon velocities is proposed and realized as well. It uses the periodicity of the fluxon moving.

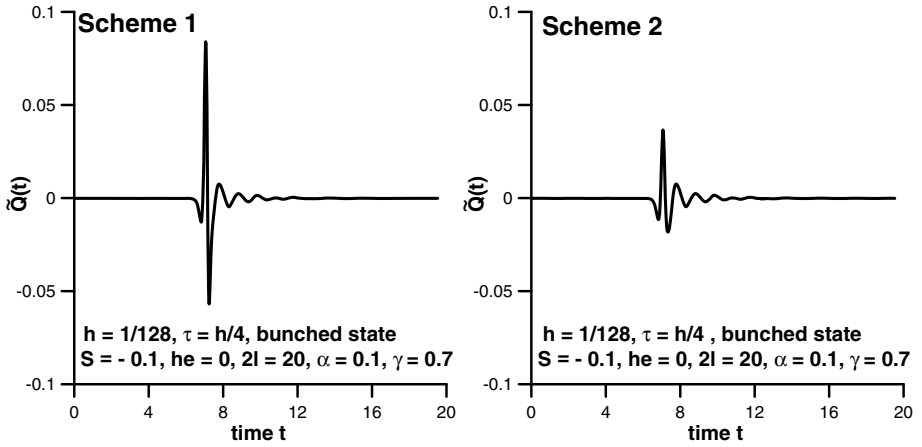


Fig. 1. Discrepancies in the integral identity: *left* - for Scheme 1, *right* - for Scheme 2

To solve numerically the static problem, corresponding to the dynamic one (1), (2), we use an iterative algorithm, based on the continuous analog of Newton's method (CANM) [9]. CANM gives a linear boundary value problem at each iteration step, which is solved numerically by means of Galerkin finite element method and quadratic finite elements. For more detailed explanation of these methods see [1].

4 Numerical Results

We show that in the case of three stacked LJJs of linear geometry the bunching of fluxons is possible. We consider two cases - one moving fluxon and two moving fluxons in each junction of the stack. We suppose that the fluxons in the exterior junctions are the same, because such systems are of physical interest. That is why the pictures for the phase gradient φ_x (which is proportional to the magnetic field) contain two graphs - one for the exterior junctions and one for the interior one. On Fig. 2 the phase gradient φ_x at some fixed time t for the case of one unbunched fluxon (left) and one bunched fluxon (right) at each junction are shown. On Fig. 3 the phase gradient φ_x at some fixed time t for the case of two unbunched fluxons (left) and two bunched fluxons (right) at each junction are shown.

The quantities which may be measured in the physical experiments, are the voltages in the individual junctions and in the whole system. So we construct the different branches of the current-voltage characteristics in mentioned above two cases. As initial data for the dynamic problem we use the stationary solutions of type (F^1, F^1, F^1) and (F^2, F^2, F^2) . The current-voltage characteristics give information about the possible behavior of the fluxon configurations when the external current γ changes.

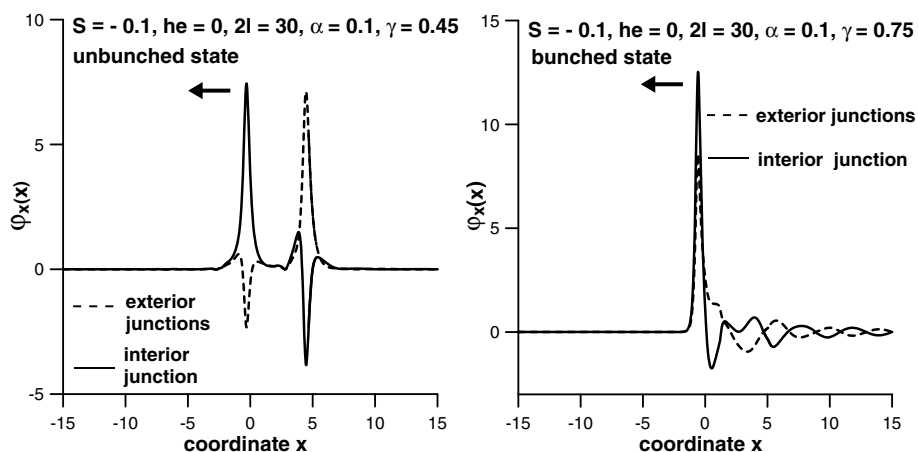


Fig. 2. One moving fluxon at each junction: *left* - unbunched case, *right* - bunched case. The arrows show the direction of moving.

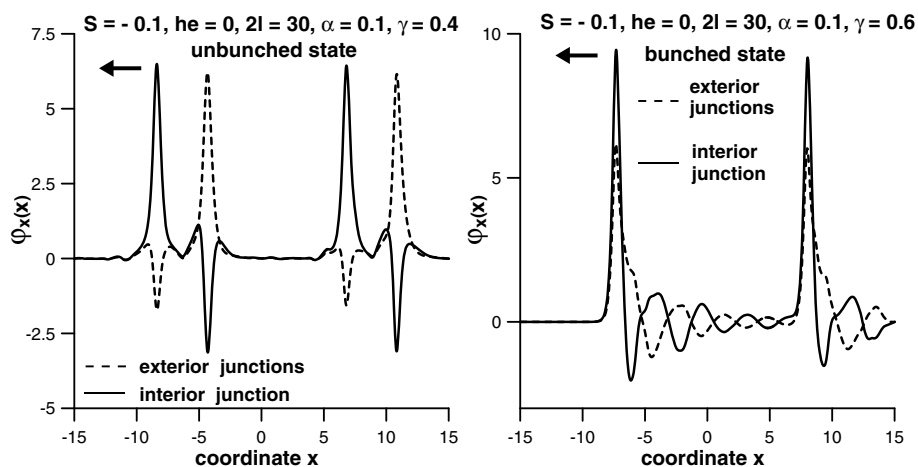


Fig. 3. Two moving fluxons at each junction: *left* - unbunched case, *right* - bunched case. The arrows show the direction of moving.

On Fig. 4 a) the branches of the current-voltage characteristics corresponding to bunched and unbunched one or two fluxons are shown for fixed, small in modulus negative value of the coupling parameter $S = -0.05$ and two different values of the parameter α ($\alpha = 0.1$ and $\alpha = 0.2$). The arrows show the states to which the solutions, corresponding to these branches, are transformed.

In the case of one fluxon in each junction (the graphs on the left of Fig. 4a)) the scenario is the following.

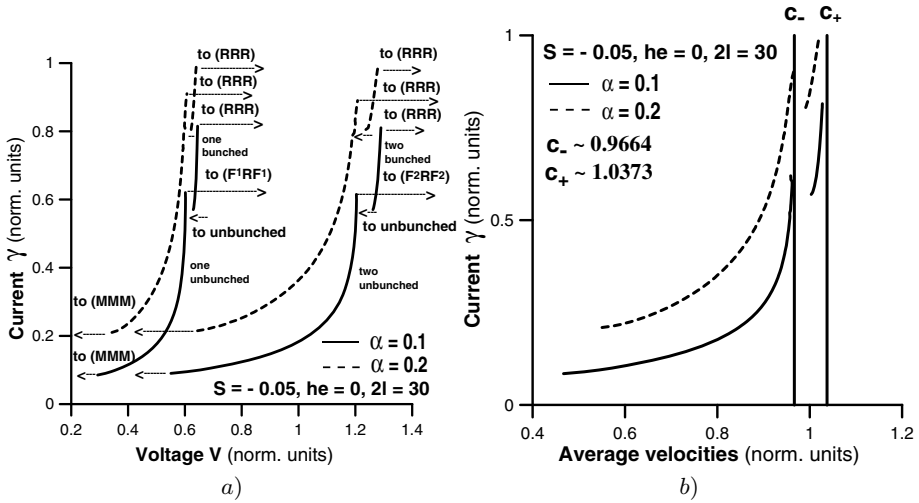


Fig. 4. a) Current-voltage characteristics and b) current-velocity characteristics for coupling constant $S = -0.05$

When γ is less than a threshold value, the bunched state does not exist. Fluxons in exterior junctions split from the fluxon in the interior junction and they propagate with different velocities (as it is shown on Fig. 2, left). In this case a movement of fluxons in opposite directions happens and this leads to complicated interaction between them.

Increasing the current γ , we find a range of values, where bunched states exist. Then the fluxons move their centers with the same velocity (as it is shown on Fig. 2, right). The emergence of oscillating tails of opposite polarities induces the three fluxons to bunch.

For a too high current the equilibrium of bunching is broken and the system switches to a new dynamic state. Let us note, that it is impossible to rebunch fluxons by increasing the current γ .

The bunched interval depends on the coupling S and the dissipation α .

For $\alpha = 0.1$ the unbunched state (F^1, F^1, F^1) transforms to (F^1, R, F^1) state for big values of γ and to (M, M, M) state for small γ . The bunched (F^1, F^1, F^1) state transforms to (R, R, R) state for big values of γ and to unbunched (F^1, F^1, F^1) state for small γ . The notion of R —(resistive) state is referred to a dynamic state, characterized by very fast and nearly uniformly rotating of the Josephson phases.

The difference in the behavior of the fluxons for $\alpha = 0.2$ is that the unbunched state transforms directly to (R, R, R) state for big values of γ .

The scenario for two fluxons in each junction (the graphs on the right of Fig. 4a)) is analogous.

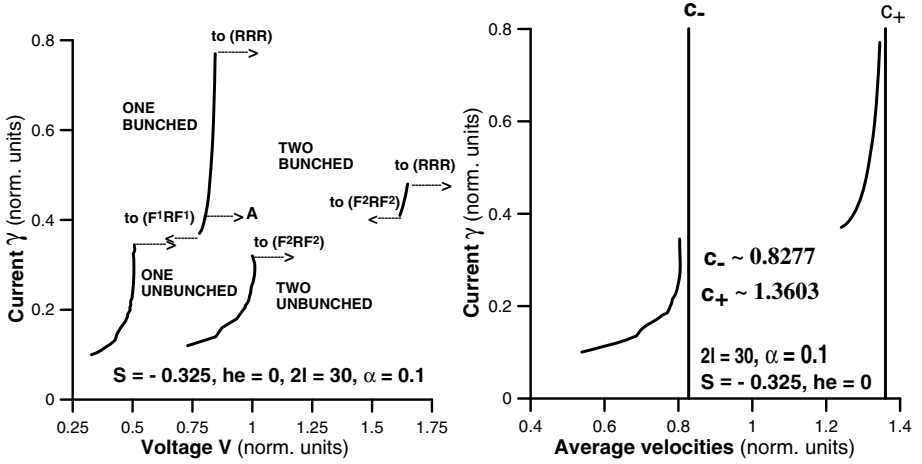


Fig. 5. Current-voltage characteristics (*left*) and current-velocity characteristics (*right*) for coupling constant $S = -0.325$

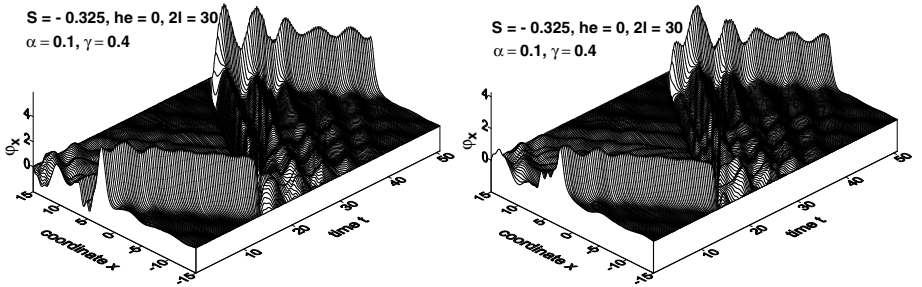


Fig. 6. Bunched state: *left* - interior junction, *right* - exterior junction

On Fig. 4b) the current-velocity characteristics for the one fluxon case are shown for the same two values of α , ($\alpha = 0.1$ and $\alpha = 0.2$). As it was expected, the fluxon bunching occurs in a velocity interval between $c_- \approx -0.9664$ and $c_+ \approx 1.0373$. For bigger in modulus negative values of the coupling parameter S ($S = -0.325$ on Fig. 5) there is no interval in γ where bunched and unbunched states exist simultaneously. In particular, it is not possible the bunch state to transforms into unbunched one, as it was for small values of S .

On Fig. 5, right, the current-velocity characteristic for the one fluxon case is shown for $\alpha = 0.1$ and $S = -0.325$. As it was expected, the fluxon bunching occurs in a velocity interval between $c_- \approx 0.8277$ and $c_+ \approx 1.3603$.

The dynamics of one bunched fluxon at each junction corresponding to point A from Fig. 5, left, is shown on Fig. 6.

5 Conclusions and Further Development

We have studied numerically the fluxon dynamics in three stacked inductively coupled LJJ's of linear geometry. The unbunched and bunched states of one and two moving fluxons in each junction are described in terms of the current-voltage and current-velocity characteristics. Different behavior of the moving fluxons is observed for small and big in modulus negative values of the coupling parameter. The detailed investigation of this dependence is forthcoming.

Acknowledgments. This work is supported by Sofia University Scientific foundation under Grant No 196/2010.

References

1. Christov, I., Dimova, S., Boyadjiev, T.: Numerical Investigation of Josephson Junction Structures. In: AIP Conference Proceedings, vol. 1186, pp. 57–68 (2009)
2. Goldobin, E., Malomed, B.A., Ustinov, A.B.: Bunching of fluxons by Cherenkov radiation in Josephson multilayers. *Phys. Rev. B* 62, 1414–1420 (2000)
3. Goldobin, E., Ustinov, A.V.: Neighbor-junction state effect on the fluxon motion in a Josephson stack. *Phys. Rev. B* 62, 1427–1432 (2000)
4. Gorria, C., Christiansen, P.L., Gaididei, I.B., Muto, V., Pedersen, N.F., Soerensen, M.P.: Fluxons dynamics in three stacked Josephson junctions. *Phys. Rev. B* 66, 172503(4) (2002)
5. Gorria, C., Christiansen, P.L., Gaididei, I.B., Muto, V., Pedersen, N.F., Soerensen, M.P.: Fluxons and their interactions in a system of three stacked Josephson junctions. *Phys. Rev. B* 68, 035415(10) (2003)
6. Gronbech-Jensen, N., Cai, D., Bishop, A.R., Lau, A.V.C., Londahl, P.S.: Bunched fluxons in coupled Josephson junctions. *Phys. Rev. B* 50, 6259–6352 (1994)
7. Kazach, G.S., Serdyukova, S.I.: Numerical Investigation of the Behaviour of Solutions of the Sine-Gordon Equation with a Singularity for Large t . *Comput. Maths. Math. Phys.* 33(3), 377–385 (1993)
8. Licharev, K.K.: *Dynamics of Josephson Junctions and Circuits*. Gordon and Breach, New York (1986)
9. Puzynin, I.V., et al.: Methods of computational physics for investigation of models of complex physical systems. *Particals & Nucley* 38 (2007)
10. Sakai, S., Bodin, P., Pedersen, N.F.: Fluxons in thin-film superconductor-insulator superlattices. *J. Appl. Phys.* 73, 2411–2418 (1993)
11. Vernik, I.V., Lazarides, N., Soerensen, M.P., Ustinov, A.V., Pedersen, N.F., Oboznov, V.A.: Soliton Bunching in Annular Josephson Junctions. *J. Appl. Phys.* 79, 7054–7060 (1996)

Global Convergence Properties of the SOR-Weierstrass Method^{*}

Vladimir Hristov², Nikolay Kyurkchiev^{1,2}, and Anton Iliev^{1,2}

¹ Faculty of Mathematics and Informatics, Paisii Hilendarski University of Plovdiv,
24, Tsar Assen Str., 4000 Plovdiv, Bulgaria

`nkyurk@uni-plovdiv.bg`, `aii@uni-plovdiv.bg`

² Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Acad. G. Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria

`vhristov@math.bas.bg`

Abstract. In this paper we give sufficient conditions for k -th approximations of the zeros of polynomial $f(x)$ under the Successive Over-Relaxation Weierstrass method (SORW) fails on the next step. This is a further improvement of the known results. Interesting numerical examples are presented.

Keywords: polynomial roots, successive over-relaxation Weierstrass method (SORW), divergent sets.

1 Introduction

Let f be a monic polynomial of degree n ,

$$f(x) := x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

with simple roots $x_i, i = 1, 2, \dots, n$. Let $x_i^k, i = 1, 2, \dots, n$, be distinct reasonably close approximations of these zeros.

The algorithm (which we refer to as SORW method) iterates as follows [9], [6], [7]:

$$x_i^{k+1} = x_i^k - h_k \frac{f(x_i^k)}{\prod_{j \neq i} (x_i^k - x_j^k)}, i = 1, \dots, n; k = 0, 1, 2, \dots, \quad (1)$$

where $h_k \in (0, 1]$ is an acceleration parameter.

SOR-like accelerations of the iterative method (1) was given in [4], [1], [3], [5], [8]. It should be noted that the method (1) has a form of prediction-correction method and converges superlinearly for $h_k < 1$.

^{*} This paper is partially supported by project IS-M-4 of Department for Scientific Research, Paisii Hilendarski University of Plovdiv.

Wang and Zhao [9] defined the acceleration parameter h_k by

$$h_k = \min_i \left(1, 0.204378 d_k \left(\sum_{i=1}^n \left| \frac{f(x_i^k)}{\prod_{j \neq i} (x_i^k - x_j^k)} \right| \right)^{-1} \right),$$

where $d_k = \min_{i \neq j} |x_i^k - x_j^k|$ which is of practical importance.

The optimal value of h_k in the sense of a guaranteed convergence is not known.

Many authors observed in practice that the method (1) is globally convergent for almost every starting point $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$ assuming that the components of $\mathbf{x}^0$ are distinct.

The following was shown in [6].

Theorem 1. *Let x_i^{k+1} be determined by (1) for $i = 1, 2, \dots, n$ and $k = 0, 1, 2, \dots$, then the following relations are valid*

$$\begin{aligned} \frac{1}{h_k} \sum_{i=1}^n x_i^{k+1} &= \left(\frac{1}{h_k} - 1 \right) \sum_{i=1}^n x_i^k - a_{n-1}, \\ \frac{1}{h_k} \sum_{i=1}^n x_i^{k+1} \sum_{j \neq i}^n x_j^k &= \left(\frac{2}{h_k} - 1 \right) \sum_{\nu < s}^n x_\nu^k x_s^k + a_{n-2}, \\ &\dots \end{aligned} \tag{2}$$

$$\frac{1}{h_k} \sum_{i=1}^n x_i^{k+1} \prod_{j \neq i}^n x_j^k = \left(\frac{n}{h_k} - 1 \right) \prod_{j=1}^n x_j^k + (-1)^n a_0$$

Theorem 2. *If the sequence of approximations x_i^k , $i = 1, \dots, n$ satisfies*

$$\begin{aligned} \left(\frac{1}{h_k} - 1 \right) \sum_{i=1}^n x_i^k - a_{n-1} &= 0, \\ \left(\frac{2}{h_k} - 1 \right) \sum_{\nu < s}^n x_\nu^k x_s^k + a_{n-2} &= 0, \\ &\dots \end{aligned} \tag{3}$$

$$\left(\frac{n}{h_k} - 1 \right) \prod_{j=1}^n x_j^k + (-1)^n a_0 = 0$$

then $x^{k+1} = (0, \dots, 0)^t$ and the predictor-correction method (1) is not defined at the $k+2$ -th approximation step, i.e. for any monic polynomial $f(x)$ of degree

n there exists a set $G_f \subset C^n$ such that the SORW method (1) starting from $x^k = x \in G_f$ does not converge to the roots of f .

This set yielding divergent starting points and obtained as the set of solutions of nonlinear systems of n equations will be called *NS-divergent set* (see, [6]).

Theorem 3. *Let f be a monic polynomial of degree n with simple roots. There is only one NS-divergent starting vector $G_f = x = (x_1, x_2, \dots, x_n) \in C^n$ for the SORW method, not counting the permutations of the component of x , and it is given by the set of roots of the algebraic polynomial*

$$F(x) = x^n - \frac{h_0 a_{n-1}}{1 - h_0} x^{n-1} - \frac{h_0 a_{n-2}}{2 - h_0} x^{n-2} - \dots - \frac{h_0 a_k}{n - k - h_0} x^k - \dots - \frac{h_0 a_0}{n - h_0}. \quad (4)$$

We observe that, in general, sets of NS-type are not the only divergent ones.

Such (4) critical initial conditions for some methods have been considered in [5], [2] and in other publications cited therein.

In this paper we give new sufficient conditions for k -th approximations of the zeros of f under which the SORW-method fails on the next step.

This is a further improvement of the known results.

2 Main Results

It should be noted that the NS-divergent set (see, Theorem 2) lead to $x_1^{k+1} = x_2^{k+1} = \dots = x_n^{k+1} = 0$ for which method (1) will fail.

We observe that the SORW method can also not be performed at the $(k+2)$ -th step if $x_i^{k+1} = x_j^{k+1}$ for some $1 \leq i < j \leq n$.

The resulting systems of equations (2) can be written in vector form as:

$$A\mathbf{x}^{k+1} = \mathbf{b},$$

where

$$A := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} \frac{1}{h_k} & \frac{1}{h_k} & \dots & \frac{1}{h_k} \\ \frac{1}{h_k} \sum_{q \neq 1} x_q^k & \frac{1}{h_k} \sum_{q \neq 2} x_q^k & \dots & \frac{1}{h_k} \sum_{q \neq n} x_q^k \\ \vdots & \vdots & & \vdots \\ \frac{1}{h_k} \prod_{q \neq 1} x_q^k & \frac{1}{h_k} \prod_{q \neq 2} x_q^k & \dots & \frac{1}{h_k} \prod_{q \neq n} x_q^k \end{pmatrix},$$

$$\det A = \frac{1}{h_k^n} \prod_{i < j}^n (x_i^k - x_j^k) \neq 0,$$

$$\mathbf{x}^{k+1} := \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \vdots \\ x_n^{k+1} \end{pmatrix},$$

$$\mathbf{b} := \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \left(\frac{1}{h_k} - 1\right) \sum_{i=1}^n x_i^k - a_{n-1} \\ \left(\frac{2}{h_k} - 1\right) \sum_{\nu < s} x_\nu^k x_s^k + a_{n-2} \\ \vdots \\ \left(\frac{n}{h_k} - 1\right) \prod_{j=1}^n x_j^k + (-1)^n a_0 \end{pmatrix},$$

We denote

$$A_j := \begin{pmatrix} a_{11} & \dots & a_{1j-1} & b_1 & a_{1j+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j-1} & b_2 & a_{2j+1} & \dots & a_{2n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj-1} & b_n & a_{nj+1} & \dots & a_{nn} \end{pmatrix},$$

$$\Delta_{sj} := \begin{vmatrix} a_{11} & \dots & a_{1j-1} & a_{1j+1} & \dots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{s-11} & \dots & a_{s-1j-1} & a_{s-1j+1} & \dots & a_{s-1n} \\ a_{s+11} & \dots & a_{s+1j-1} & a_{s+1j+1} & \dots & a_{s+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj-1} & a_{nj+1} & \dots & a_{nn} \end{vmatrix}.$$

Clearly,

$$\det A_j = \sum_{s=1}^n (-1)^{j+s} b_s \Delta_{sj}, \quad j = 1, 2, \dots, n.$$

We have the following theorem.

Theorem 4. *Suppose that for some $1 \leq i < j \leq n$, the sequence of approximations $x_1^k, \dots, x_n^k$ satisfies the condition*

$$\det A_i - \det A_j = 0. \quad (5)$$

Then $x_i^{k+1} = x_j^{k+1}$, and thus, the $(k+2)$ -th step of the SORW-method cannot be performed.

Proof. The proof follows the ideas given in [2]. We note that, if $x_i^{k+1} = x_j^{k+1}$, then by Cramer's formula

$$\frac{\det A_i}{\det A} = \frac{\det A_j}{\det A}.$$

$\det A \neq 0$, and we arrive at the formulae (5), which completes the proof of Theorem 4.

The set D_f of the non-attractive starting points is the set of points satisfying equations (5).

3 Examples

1. For illustration, we consider non-attractive set D_f in the example of the equation

$$f(x) = x^3 + x^2 + x + 1 = 0. \quad (6)$$

The non-attractive set D_f , for instance, if $i = 1, j = 2$, is given by (see, (5))

$$\begin{aligned} D_f : &= z^2(x^2 + y^2 - 2xy) \\ &- z((1-h)x^3 + (1-h)y^3 - x^2y - xy^2 - hx^2 - hy^2 - hx - hy - 2h) \\ &+ (1-h)x^3y - 2x^2y^2 + (1-h)xy^3 \\ &- hx^2y - hxy^2 - 2hxy - hx - hy = 0, \end{aligned}$$

where

$$z = x_3^k; \quad x = x_1^k; \quad y = x_2^k; \quad h = h_k$$

and displayed in Figure 1 for $h = 0.8$.

For the classical Weierstrass method see, Figure 2.

It is clear that the choice of a small h_k is not advisable.

In some cases the choice of $h_k < 1$ can enable a good start of the method (1), which depends on the structure of a polynomial and the distribution of initial approximations [7].

The non-attractive set D_f for $h = 0.2$ displayed in Figure 3.

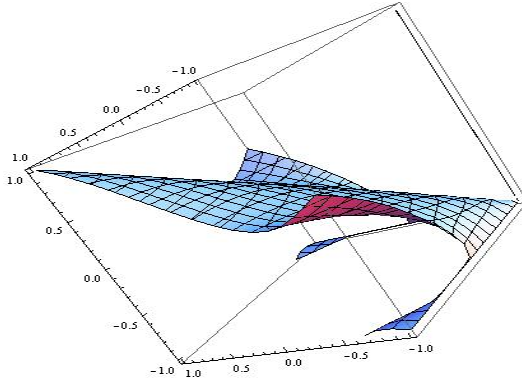


Fig. 1.

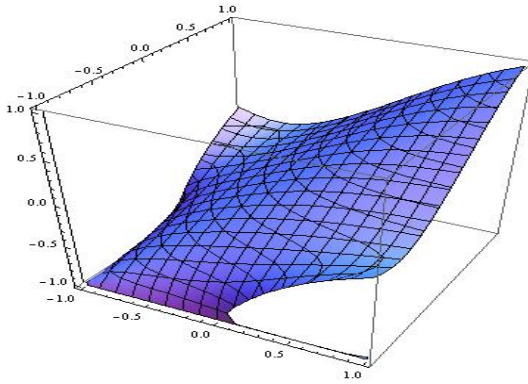


Fig. 2.

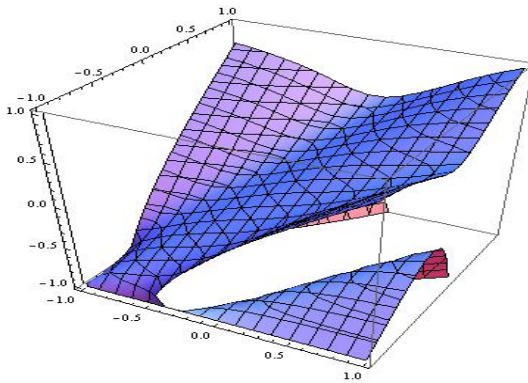


Fig. 3.

2. For the equation

$$f(x) = x^3 + 1 = 0. \quad (7)$$

the non-attractive set D_f , for instance, if $i = 1$, $j = 2$, is given by

$$D_f := z^2(x^2 + y^2 - 2xy) - z((1-h)x^3 + (1-h)y^3 - x^2y - xy^2 - 2h) + \\ (1-h)x^3y - 2x^2y^2 + (1-h)xy^3 - hx - hy = 0,$$

and displayed in Figure 4 for $h = 0.9$.

We observe that the convergence of the SORW method is faster if h_k is closer to 1.

We found that the method (1) satisfied the stopping criterion

$$\max_{1 \leq i \leq n} |f(x_i^k)| < \varepsilon = 10^{-6}; \quad k = 0, 1, \dots$$

and minimal number of iterations is obtained for $h \in [0.85, 1]$.

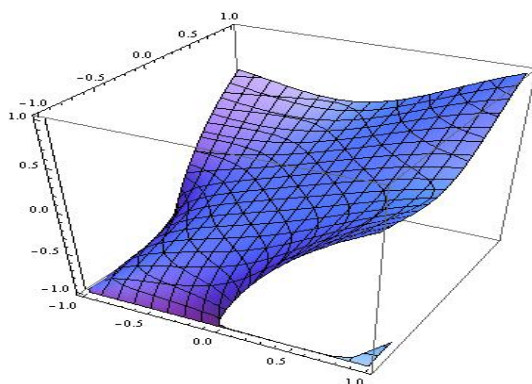


Fig. 4.

References

1. Atanassova, L., Kyurkchiev, N., Yamamoto, T.: Methods for computing all roots of a polynomial simultaneously - known results and open problems. *Computing [Suppl.]* 16, 23–35 (2002)
2. Hristov, V., Kyurkchiev, N.: A note on the globally convergent properties of the Weierstrass-Dochev method. In: Boyanov, B. (ed.) *Approximation Theory. A volume dedicated to Bl. Sendov*, DARBA, Sofia, pp. 231–240 (2002)
3. Kyurkchiev, N.: A note on the convergence of the SOR-like Weierstrass method. *Computing [Suppl.]* 16, 143–149 (2002)
4. Kanno, S., Kyurkchiev, N., Yamamoto, T.: On some methods for the simultaneous determination of polynomial zeros. *Japan J. Indust. Appl. Math.* 13, 267–288 (1996)
5. Kyurkchiev, N.: Initial approximations and root finding methods. In: *Mathematical Research*, vol. 104. Wiley-VCH Verlag Berlin GmbH, Berlin (1998)
6. Kyurkchiev, N., Petkovic, M.: On the behaviour of approximations of the SOR Weierstrass method. *Comput. Math. with Appl.* 32, 117–121 (1996)

7. Petkovic, M., Kyurkchiev, N.: A note on the convergence of the Weierstrass SOR method for polynomial roots. *J. Comput. Appl. Math.* 80, 163–168 (1997)
8. Petkovic, M., Herceg, D., Ilic, S.: Point estimation theory and its applications. Institute of Mathematics. Novi. Sad. (1997)
9. Zhao, F., Wang, D.: The theory of Smale’s point estimation and its applications. *J. Comput. Appl. Math.* 60, 253–269 (1995)

Numerical Solution of a Nonlinear Evolution Equation for the Risk Preference

Naoyuki Ishimura¹, Miglena N. Koleva², and Lubin G. Vulkov²

¹ Graduate School of Economics, Hitotsubashi University, Tokyo 186-8601, Japan
ishimura@econ.hit-u.ac.jp

² Faculty of Natural Science and Education, University of Rouse,
8 Studentska str., Rouse 7017, Bulgaria
{mkoleva,lvulkov}@uni-ruse.bg

Abstract. A singular nonlinear partial differential equation (PDE) for the risk preference was derived by the first author in previous publications. The PDE is related to the Arrow-Pratt coefficient of relative risk aversion. In the present paper, we develop a Rothe-Bellman & Kalaba quasilinearization method on quasi-uniform space mesh to numerically investigate such PDE. Numerical experiments are discussed.

1 Introduction

The optimal behaviors in economics environment have been an intensive subject for researches. Various models have been based on stochastic control framework and the analysis is often reduced to the treatment of the Hamilton-Jacobi-Bellman (HJB) equation for the value function. The HJB equations, however, are fully nonlinear and hard to solve even in the weak viscosity solution sense.

In this article we deal with a singular nonlinear partial differential equation (PDEs) which is derived from the HJB equation for the value function in the optimal investment problem. The derived PDE is quasilinear and the unknown quantity is related to the Arrow-Pratt coefficient of relative risk aversion [10] with respect to the optimal value function. We remark that in our previous study [1][5], the PDE is related to the Arrow-Pratt coefficient of the “absolute” risk aversion, while the current equation is on the “relative” risk aversion.

We begin with recalling our model. Suppose that the wealth X_t at time t (≥ 0) of the company is subject to a fluctuating process, and the company wants to invest in one risky stock, whose price P_t is governed by the stochastic differential equation $dP_t = P_t(\mu dt + \sigma dW_t^{(1)})$, where μ, σ are constants and $\{W_t^{(1)}\}_{t \geq 0}$ is a standard Brownian motion. The fluctuating process, denoted by Y_t , is assumed to be $dY_t = \alpha dt + \beta dW_t^{(2)}$, where α, β ($\beta > 0$) are constants and $\{W_t^{(2)}\}_{t \geq 0}$ is another standard Brownian motion. These two Brownian motions are allowed to be correlated with the correlation coefficient ρ ($0 \leq |\rho| < 1$). The investment policy $f = \{f_t\}_{0 \leq t \leq T}$ (T stands for the maturity) of the company is a suitable admissible adapted control process. The process of X_t^f is then assumed to be $\frac{dX_t^f}{X_t^f} = f_t \frac{dP_t}{P_t} + dY_t = (f_t \mu + \alpha)dt + f_t \sigma dW_t^{(1)} + \beta dW_t^{(2)}$, $X_0^f = x \in \mathbf{R}$.

Suppose that the company aims to maximize the utility $u(x)$ ($u' > 0$ and $u'' < 0$ is assumed) from his terminal wealth. Let

$$V(x, t) := \sup_f E[u(X_T^f) | X_t^f = x]. \quad (1)$$

Then the Hamilton-Jacobi-Bellman equation for the value function (1) becomes

$$\sup_f \{\mathcal{A}^f V(x, t)\} = 0, \quad V(x, T) = u(x), \quad (2)$$

where $(\mathcal{A}^f g)(x, t) := \frac{\partial g}{\partial t} + (f\mu + \alpha)x \frac{\partial g}{\partial x} + \frac{1}{2}(f^2\sigma^2 + \beta^2 + 2\beta\sigma\rho f)x^2 \frac{\partial^2 g}{\partial x^2}$.

Suppose that (2) has a classical solution V with $\partial V/\partial x > 0$, $\partial^2 V/\partial x^2 < 0$. We then discover that the optimal policy is $\{f_t^*\}_{0 \leq t \leq T}$ $f_t^* = -\frac{\mu}{\sigma^2} \frac{\partial V/\partial x}{\partial^2 V/\partial x^2} - \frac{\beta\rho}{\sigma}$. Plugging this back into (2) we obtain

$$\begin{aligned} \frac{\partial V}{\partial t} + \left(\alpha - \frac{\beta\rho\mu}{\sigma}\right)x \frac{\partial V}{\partial x} - \frac{\mu^2}{2\sigma^2} \frac{(\partial V/\partial x)^2}{\partial^2 V/\partial x^2} + \frac{1}{2}\beta^2(1 - \rho^2)x^2 \frac{\partial^2 V}{\partial x^2} &= 0, \quad 0 < t < T, \\ V(T, x) &= u(x). \end{aligned}$$

Let $\tau := 2(1 - \rho^2)^{-1}\beta^{-2}(T - t)$ and put $W(x, \tau) = V(x, t(\tau))$, we find that

$$\frac{\partial W}{\partial \tau} = x^2 \frac{\partial^2 W}{\partial x^2} - a^2 \frac{(\partial W/\partial x)^2}{\partial^2 W/\partial x^2} - bx \frac{\partial W}{\partial x}, \quad W(x, 0) = u(x), \quad (3)$$

where $a^2 := \frac{\mu^2}{(1-\rho^2)\beta^2\sigma^2}$, $b := \frac{2(\rho\mu\beta - \alpha\sigma)}{(1-\rho^2)\beta^2\sigma}$. The equation (3) is fully nonlinear, see [7]. We define $\tilde{r}(x, \tau) := -\frac{x\partial^2 W/\partial x^2}{\partial W/\partial x} = -x \frac{\partial}{\partial x} \log \left| \frac{\partial W}{\partial x}(x, \tau) \right|$, which extends the Arrow-Pratt coefficient of relative risk aversion for the utility function. We note that similar transformation $-W_x/W_{xx}$ is considered in [12]. Making a change of variables $x = e^y$ ($y = \log x$) and putting $r(y, \tau) = \tilde{r}(x(y), \tau)$ (see [8]); we infer that

$$\begin{aligned} \frac{\partial r}{\partial \tau} &= \left(\frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial y} \right) \left(r - \frac{a^2}{r} \right) - (2r + b) \frac{\partial r}{\partial y} \\ &=: \mathcal{L} \left(r, \frac{\partial r}{\partial y}, \frac{\partial^2 r}{\partial y^2} \right) \quad \text{for } -\infty < y < \infty, \quad 0 < \tau < T. \end{aligned} \quad (4)$$

The rest of the paper is organized as follows. In Section 2 a travelling wave solution is constructed. The numerical method is described in Section 3. Numerical results using as a test solution the travelling wave solution are given in Section 4. Finally, some conclusions put an end to the paper.

2 Travelling Wave Solution

For standard risk averse investor, the coefficient of relative risk aversion is expected to be non-increasing [9]. Observing that every constant function verifies

the equation (4), we wish to seek a travelling wave solution $r = r(y - v\tau)$ (the wave speed $v \in \mathbf{R}$ should be determined later) with the property

$$\begin{aligned} r'(y) &< 0 & \text{for } -\infty < y < \infty, \\ r(y) &\rightarrow r_- \text{ as } y \rightarrow -\infty, \quad \text{and} \quad r(y) \rightarrow r_+ \text{ as } y \rightarrow \infty, \end{aligned} \quad (5)$$

where $r_- > r_+ > 0$ are prescribed constants. Putting $r(y, \tau) = r(y - \tau v)$ into (4), we derive an ordinary differential equation for $r = r(y)$. Integrating once, we obtain

$$\left(1 - \frac{a^2}{r}\right)' + r - \frac{a^2}{r} - r^2 - br + vr = C. \quad (6)$$

Here C denotes a constant, and from the boundary condition (5) we deduce that

$$C = r_- r_+ - \frac{r_- + r_+}{r_- r_+} a^2, \quad v = r_- + r_+ - \frac{a^2}{r_- r_+} + b - 1.$$

The equation (6) can be written in the separable form $\frac{r^2 + a^2}{r(r^3 + (b-v-1)r^2 + Cr + a^2)} dr = dy$. We define $f(r) := r^3 - (v+1-b)r^2 + Cr + a^2$ and with precise analysis of the criterion $f(r_-) = f(r_+) = 0$ as well as tedious calculations we obtained that: *for any $r_- > r_+ > 0$ satisfying $r_- r_+ (r_- + r_+) > a^2$, there exists a travelling wave solution $r = r(y - v\tau)$ with $v = r_- + r_+ - r_-^{-1} r_+^{-1} a^2 + b - 1$ to (4) such that*

$$r'(y) < 0 \quad \text{for } -\infty < y < \infty \quad \text{and} \quad r(y) \rightarrow r_{\pm} \text{ as } y \rightarrow \pm\infty, \text{ respectively.} \quad (7)$$

3 Numerical Method

In this section, we undertake the numerical study on the equation (4).

3.1 Semi-discretization in Time

We apply Rothe's method, which is commonly used as a numerical approximation. This method corresponds to doing a backward Euler approximation and is also known sometimes as the method of lines. Let divide the interval $[0, T]$ into n subintervals of length $\Delta\tau$, to obtain the mesh

$$\overline{\omega}_\tau = \{\tau_j = j\Delta\tau, j = 0, \dots, n, \tau_0 = 0, \tau_n = T\}.$$

For each $\tau = \tau_j$ we approximate the unknown function $r(y, \tau)$ by $r^j(y)$ and the derivative $\partial r / \partial \tau$ by the difference quotient

$$\frac{\partial r}{\partial \tau}(y, j\Delta\tau) \approx (r^j(y) - r^{j-1}(y)) / \Delta\tau, \quad j = 1, \dots, n,$$

where r^{j-1} is the solution at previous time level. Starting by $r^0(y) := r(y, 0)$, the functions r^j , $j = 1, \dots, n$ are determined subsequently as solutions of ODEs

$$\mathcal{L}\left(r^j, \frac{dr^j}{dy}, \frac{d^2 r^j}{dy^2}\right) - \frac{r^j}{\Delta\tau} = -\frac{r^{j-1}}{\Delta\tau}. \quad (8)$$

Having obtained $r^1(y), r^2(y), \dots, r^n(y)$, the so-called Rothe's function $r_n(y, \tau)$ is defined in the whole region by

$$r_n(y, \tau) = r^{j-1}(y) + \frac{r^j(y) - r^{j-1}(y)}{\Delta\tau}(\tau - \tau_{j-1}), \quad \tau_{j-1} < \tau < \tau_j, \quad j = 1, \dots, n,$$

which assume the values r^j at every $\tau = \tau_j$. By refining the original division ($\Delta\tau_s$, $s = 1, \dots, \Delta\tau_s \rightarrow 0$, $s \rightarrow \infty$), we obtain the sequence $r_{n_s}(\tau)$ for corresponding Rothe's function, which can be expected to converge (in an appropriate space) to the solution u (in an appropriate sense) of the given problem.

Next, for solving the equation (4) approximately, we consider the second order θ -Rothe's difference scheme, $\theta \in [0, 1]$

$$\theta \mathcal{L} \left(r^j, \frac{dr^j}{dy}, \frac{d^2r^j}{dy^2} \right) + (1 - \theta) \mathcal{L} \left(r^{j-1}, \frac{dr^{j-1}}{dy}, \frac{d^2r^{j-1}}{dy^2} \right) - \frac{r^j}{\Delta\tau} = -\frac{r^{j-1}}{\Delta\tau}. \quad (9)$$

Blank and Smith have studied the convergence of Rothe's method for fully non-linear parabolic equations, see [4]. They show that the Rothe solutions are Lipschitz in time, Hölder in space and they solve the equation (8) in the viscosity sense with *rate of convergence* $\mathcal{O}(\Delta\tau)$. We can apply this results for our problem. *Such questions are not in the focus* of the present paper and further we will assume that the *solutions exist and have the required from the numerical method smoothness in time and space*.

3.2 Quasilinearization

We employ the quasilinearization method (QLM) of Bellman and Kalaba [3], for which iterations are constructed to yield rapidconvergence and often monotonicity. We rewrite the equations (8), $\theta = 1$ and (9) in the form

$$\mathcal{E} \left(r^j, \frac{dr^j}{dy}, \frac{d^2r^j}{dy^2} \right) = F^{j-1}, \quad \text{where} \quad (10)$$

$$F^j := \frac{r^j}{\Delta\tau} - (1 - \theta) \mathcal{L} \left(r^j, \frac{dr^j}{dy}, \frac{d^2r^j}{dy^2} \right), \quad \mathcal{E} := \theta \mathcal{L} \left(r^j, \frac{dr^j}{dy}, \frac{d^2r^j}{dy^2} \right) + \frac{r^j}{\Delta\tau}.$$

The QLM prescription [3] determines the $(k + 1)^{\text{st}}$ iterative approximation $r^{(k+1)}(y)$ to the solution of (10) as one of the linear differential equation

$$\begin{aligned} \frac{\partial \mathcal{E}^{(k)}}{\partial r}(y) \delta r^{(k+1)}(y) + \frac{\partial \mathcal{E}^{(k)}}{\partial r_y}(y) \frac{d}{dy} [\delta r^{(k+1)}(y)] + \frac{\partial \mathcal{E}^{(k)}}{\partial r_{yy}}(y) \frac{d^2}{dy^2} [\delta r^{(k+1)}(y)] \\ = -\mathcal{E}^{(k)} + F^{j-1}, \end{aligned} \quad (11)$$

where $\delta r^{(k+1)}(y) = r^{(k+1)}(y) - r^{(k)}(y)$ and $\mathcal{E}^{(k)}$ is $\mathcal{E}$ computed on $(k)^{\text{th}}$ iteration, or written in details the coefficients in (11) are

$$\begin{aligned}\frac{\partial \mathcal{E}^{(k)}}{\partial r}(y) &= \frac{1}{\Delta \tau} + 2\theta \left(\frac{d^2 r^{(k)}}{dy^2} \frac{a^2}{(r^{(k)})^3} + \frac{dr^{(k)}}{dy} \left(-\frac{dr^{(k)}}{dy} \frac{3a^2}{(r^{(k)})^4} + \frac{a^2}{(r^{(k)})^3} + 1 \right) \right), \\ \frac{\partial \mathcal{E}^{(k)}}{\partial r_y}(y) &= \theta \left(\frac{dr^{(k)}}{dy} \frac{4a^2}{(r^{(k)})^3} - \frac{a^2}{(r^{(k)})^2} + 2r^{(k)} + b - 1 \right), \\ \frac{\partial \mathcal{E}^{(k)}}{\partial r_{yy}}(y) &= -\theta \left(\frac{a^2}{(r^{(k)})^2} + 1 \right).\end{aligned}$$

The zero approximation $r^{(0)}$ is chosen from mathematical and financial mathematics considerations. *The QLM procedure yields a quadratic and often monotone convergence to the solution of problem (8) or (9), see [6,7].*

3.3 Meshes and Full Discretization

Here we consider the problem (4) with conditions (7). In order to derive an appropriate approximation of the model problem, a natural approach is using a *quasiuniform mesh* (QUM), see [2]. The obtained discretization involves original “boundary” conditions.

Let $y(\xi)$, $\xi \in [0, 1]$, $y \in [\alpha, \beta]$ is strong monotone sufficiently smooth function. Then the mesh $w_N = \{y_i = y(\frac{i}{N}), 0 \leq i \leq N\}$ in $[\alpha, \beta]$ we call *quasi-uniform*, [2]. We shall implement our problem the QUM mesh ω_h , see Figure 1.

$$\begin{aligned}\omega_h &= \begin{cases} y(\xi) = y^-(\xi), & y \leq 0, \\ y(\xi) = y^+(\xi), & y \geq 0 \end{cases}, \quad m_1 + m_2 = N, \quad y^-(1) = y^+(0) = 0, \\ y^-(\xi) &= c_1 \ln(\xi), \quad h_{m_1-1}^- = y_{m_1}^- - y_{m_1-1}^- \simeq \frac{c_1}{m_1}, \quad y_1^- = c_1 \ln(m_1) \quad (12) \\ y^+(\xi) &= -c_2 \ln(1 - \xi), \quad h_0^+ = y_1^+ - y_0^+ \simeq \frac{c_2}{m_2}, \quad y_{m_2-1}^+ = c_2 \ln(m_2), \quad (13)\end{aligned}$$

where $c_1 > 0$ and $c_2 > 0$ are controlling parameters. The choice of c_1 and c_2 are coming from the fact that the half of intervals are in domain with length $\sim c_1 + c_2$. The first interval of (12): $[y_0^-, y_1^-]$ is infinite, but the point $y_{1/2}^-$ is finite, since the non-integer nodes are given by $y_{i+\alpha}^- = y^-(\frac{i+\alpha}{m_1})$, $|\alpha| < 1$. The same is for $y^+(\xi)$: the last interval of (13): $[y_{m_2-1}^+, y_{m_2}^+]$, is infinite, but the point $y_{m_2-1/2}^+$ is finite, since the non-integer nodes are given by $y_{i+\alpha}^+ = y^+(\frac{i+\alpha}{m_2})$, $|\alpha| < 1$. Therefore, the QUM transforms the infinite domain into finite number of intervals and places the original boundary condition directly on infinity.

On the base of the finite difference $\left(\frac{dr}{dy}\right)_{i+1/2} \approx (r_{i+1} - r_i)/[2(y_{i+3/4} - y_{i+1/4})]$, we derive the following derivative approximations at integer grid nodes. We note that the formulas contain $r(-\infty, t) = r_-$ and $r(+\infty, t) = r_+$, but not $y_0^- = -\infty$

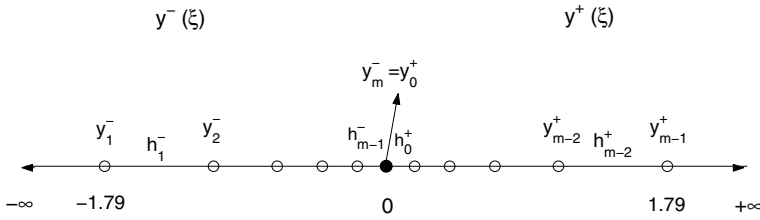


Fig. 1. QUM, $c_1 = c_2 = 1$, $m = m_1 = m_2 = 6$

and $y_N^+ = \infty$.

$$\left(\frac{dr}{dy}\right)_i \approx \frac{1}{2} \left[\left(\frac{dr}{dy}\right)_{i-1/2} + \left(\frac{dr}{dy}\right)_{i+1/2} \right],$$

$$\left(\frac{d^2r}{dy^2}\right)_i \approx \frac{1}{y_{i+1/2} - y_{i-1/2}} \left[\left(\frac{dr}{dy}\right)_{i+1/2} - \left(\frac{dr}{dy}\right)_{i-1/2} \right].$$

The truncation errors are of order $\mathcal{O}(N^{-2})$.

At point $y = 0$, where the two meshes $y^-(\xi)$ and $y^+(\xi)$ overlaps, we use the standard central first and second order derivative approximation [11] on a nonuniform (uniform) mesh, if $c_1 \neq c_2$ and $m_1 \neq m_2$ ($c_1 = c_2$ and $m_1 = m_2$).

4 Numerical Experiments

In this section we present some results for numerical solution, obtained by QLM on QUM. In the first group of experiments, we deal with exact solution in order to demonstrate second order rate of convergence in space and first ($\theta = 1$) or second ($\theta = 0.5$) order with respect to time variable. The errors $E_i = r_{exact}(y_i, T) - r_{numer}(y_i, T)$, $i = 1, \dots, N - 1$ are given in maximal and L_2 discrete norms

$$\|E^N\|_\infty = \max_{1 \leq i \leq N-1} |E_i|, \quad \|E^N\|_2 = \left[\sum_{i=1}^{N-1} (y_{i+1/2} - y_{i-1/2}) E_i^2 \right]^{1/2}$$

and the convergence rate is calculated using double mesh principle

$$CR_\infty = \log 2 (\|E^N\|_\infty / \|E^{2N}\|_\infty), \quad CR_2 = \log 2 (\|E^N\|_2 / \|E^{2N}\|_2).$$

In the next group of experiments the original solution is computed.

The mesh parameters are $m = m_1 = m_2$, $c_1 = c_2 = 1$ and $a = b = 1$, $T = 1$ for all computations. The QLM iteration procedure continue until the maximum difference between two subsequent iterations is less than 10^{-12} .

Example 1 (Exact solution). In the right hand side of (4) we add an appropriate function $f(y, t)$, such that $r(y, t) = e^{-t} \operatorname{erfc}(y) + 2$ is the exact solution of the obtained equation, associated with conditions (7). Thus, $r_- > r_+ > 0$ and $r'(y) < 0$. The ratios $\frac{\tau}{h^2} = 5$ for $\theta = 1$ and $\frac{\tau}{h} = 5$ for $\theta = 0.5$ are fixed, $h = \frac{1}{N}$.

Table 1. Error and convergence rate in maximal and L_2 discrete norms, Example 1

N	$\theta = 1$				$\theta = 0.5$			
	$\ E^N\ _\infty$	CR_∞	$\ E^N\ _2$	CR_2	$\ E^N\ _\infty$	CR_∞	$\ E^N\ _2$	CR_2
24	2.6612e-2		5.1486e-2		8.9170e-3		1.1935e-2	
48	6.8733e-3	1.9530	1.3471e-2	1.9343	2.4166e-3	1.8836	4.5917e-3	1.8041
96	1.8817e-3	1.8690	3.7508e-3	1.8446	6.6000e-4	1.8724	1.2619e-3	1.8634
192	4.9469e-4	1.9274	4.0004e-3	1.9066	1.7534e-4	1.8926	3.7108e-4	1.8874

A linear interpolation in time is used, in order to obtain the results at $T = 1$. In Table 1 we give error and convergence rate of the numerical solution, computed with QLM on QUM for $\theta = 1$ and $\theta = 0.5$, respectively. The results show that for $\theta = 1$ the accuracy is $\mathcal{O}(\tau + N^{-2})$ and for $\theta = 0.5$ the accuracy is $\mathcal{O}(\tau^2 + N^{-2})$ both in maximal and L_2 norm. The intervals, close to the infinite "boundaries" are long, which explains the results in L_2 norms.

Example 2 (Original problem). We compute the problem (4),(7), $r_- = 3$, $r_+ = 1$ with QLM on QUM. In order to start the QLM procedure we chose $r^{(0)}$ to be the solution of the ordinary differential equation (6), approximated in the same manner as equation (4). To compute the solution of problem (6),(7) we start with the initial guess $r^{(0)} = \text{erfc}(y/2) + 1$, which satisfies the conditions on the infinity "boundaries". In Figure 2 we plot *initial solution* (solution of the problem (6),(7)) and evolution graphics of the numerical solution of the problem (4),(7), $N = 96$, $\theta = 1$, $\tau = 0.05$ from $t = 0.05$ to $t = 1$. Thus the travelling wave solution is verified. Also, we clearly see that for numerical solution of the model problem (4),(7), it is not a good approach to use a large enough finite interval and to impose the boundary conditions as on the infinity, especially for long time computations.

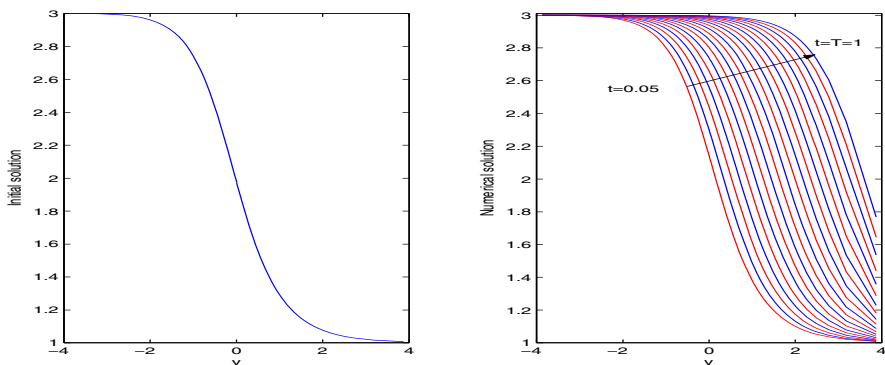


Fig. 2. Initial solution and travelling wave solution, Example 2

5 Discussions

We have introduced a singular quasilinear parabolic equation for the risk preference. Unknown function is related to the coefficient of relative risk aversion with respect to the value function in the optimal investment problem. We establish the existence of travelling wave solutions, which is welcome in the standpoint of financial economics. Concerning the numerical solution of the Cauchy problem for proposed PDEs, a combined Rothe-Bellman & Kalaba quasilinearization method is employed. Finally, we discuss numerical experiments to investigate the solution behavior and to test the accuracy of the numerical method on the exact travelling wave solution. The efficiency of the proposed numerical method is demonstrated.

Acknowledgements. The first author is supported in part by the JSPS grant (No.21540117), and the other two authors are supported by the Bulgarian National Fund of Science under Projects Sk-Bg-203 and ID_09_0186.

Authors would like to thank Referees for the helpful comments and suggestions.

References

1. Abe, R., Ishimura, N.: Existence of solutions for the nonlinear partial differential equation arising in the optimal investment problem, *Proc. Japan Acad. Ser. A* 84, 11–14 (2008)
2. Alshina, E., Kalitkin, N., Panchenko, S.: Numerical solution of boundary value problem in unlimited area. *Math. Modelling* 14(11), 10–22 (2002) (in Russian)
3. Bellman, R., Kalaba, R.: *Quasilinearization and nonlinear boundary-value problems*. Elsevier Publishing Company, New York (1965)
4. Blank, I., Smith, P.: Convergence of Rothe's method for fully nonlinear parabolic equations. *J. of Geom. Analysis* 15(3), 363–373 (2005)
5. Ishimura, N., Murao, K.: Nonlinear evolution equations for the risk preference in the optimal investment problem. Paper Presented at AsianFA/NFA 2008 International Conference in Yokohama, <http://fs.ics.hit-u.ac.jp/nfa-net/>
6. Koleva, M.N., Vulkov, L.G.: Two-grid quasilinearization approach to ODEs with applications to model problems in physics and mechanics. *Comp. Phys. Commun.* 181(3), 663–670 (2010)
7. Koleva, M.N., Vulkov, L.G.: Quasilinearization numerical scheme for fully nonlinear parabolic problems with applications in models of mathematical finance (submitted)
8. Macová, Z., Ševčovič, D.: Weakly nonlinear analysis of the Hamilton-Jacobi-Bellman equation arising from pension saving management. *Int. J. Numer. Anal. Model.* 7(4), 619–693 (2010)
9. Mas-Collel, A., Michael, D.W., Green, J.R.: *Microeconomic Theory*. Oxford University Press, Oxford (1995)
10. Pratt, J.W.: Risk aversion in the small and in the large. *Econometrica* 32, 122–136 (1964)
11. Samarskii, A.A.: *The Theory of Difference Schemes*. Marcel Dekker Inc., New York (2001)
12. Songzhe, L.: Existence of solutions to initial value problem for a parabolic Monge-Ampère equation and application. *Nonl. Anal.* 65, 59–78 (2006)

A Numerical Approach for the American Call Option Pricing Model

Juri D. Kandilarov¹ and Radoslav L. Valkov²

¹ Department of Mathematics, University of Rousse
ukandilarov@ami.uni-ruse.bg

² Faculty of Mathematics and Informatics, University of Sofia
rvalkov@fmi.uni-sofia.bg

Abstract. We present a numerical approach of the free boundary problem for the Black-Scholes equation for pricing the American call option on stocks paying a continuous dividend. A fixed domain transformation of the free boundary problem into a parabolic equation defined on a fixed spatial domain is performed. As a result a nonlinear time-dependent term is involved in the resulting equation. Two iterative numerical algorithms are proposed. Computational experiments, confirming the accuracy of the algorithms are discussed.

1 Introduction

Analytical solutions of Black-Scholes model option problems are seldom available and hence such derivatives must be priced by numerical techniques. The problem of solving the American option problem numerically during the last decade has been subject for intensive research [1, 2, 4–6, 9, 12–17]. Elementary introduction to this topic can be found in [5]. Qualitative and quantitative comparison of various analytical and numerical approximation methods of calculation a position of the early exercise boundary of the American put option paying zero dividends is given in [14]. An improvement of Han and Wu’s algorithm [4] is described in [15]. In this paper we introduce two front-fixing numerical algorithms for solving free and moving boundary value problem, formulated in [5, 12, 14, 16].

The front-fixing method has been applied successfully to a wide range problems arising in physics and engineering, cf. [3, 7, 8] and references therein. The basic idea is to remove the moving boundary by a transformation of the involved variables. In the survey chapter [12] a transformation technique that can be used in analysis and numerical computation of the early exercise boundary for an American style of vanilla options that can be modeled by class of generalized Black-Scholes equations is presented.

In this paper we show how this technique can be extended for the American call-option problem. Furthermore, we present an implicit and an explicit algorithms for solving the resulting nonlinear difference systems. The outline of the paper is as follows. In the next section we define the front-fixing method for the Black-Scholes model for American call option. In Section 3 we derive the finite difference schemes and the associate iterative numerical algorithms. Finally,

in Section 4 several numerical experiments illustrating the performance of our algorithms are discussed.

2 The Free Boundary Problem

American call option is the following PDE:

$$\frac{\partial V}{\partial t} + (r - D)S \frac{\partial V}{\partial S} + \frac{\sigma^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} - rV = 0, \quad 0 < t < T, \quad 0 < S < S_f(t), \quad (1)$$

$$V(0, t) = 0, \quad V(S_f(t), t) = S_f(t) - E,$$

$$\frac{\partial V}{\partial S}(S_f(t), t) = 1, \quad V(S, T) = \max(S - E, 0), \quad (2)$$

defined on a time-dependent domain $S \in (0, S_f(t))$, where $t \in (0, T)$. Here $S > 0$ is the stock price, $E > 0$ is the exercise price, $r > 0$ is the risk-free rate, $D > 0$ is the continuous stock dividend rate and $\sigma > 0$ is the volatility of the underlying stock process.

In this paper we restrict our attention to the case when $r > D > 0$. It is well known that for $r > D > 0$ the free boundary $\rho(\tau) = S_f(T - \tau)$ starts at $\rho(0) = rE/D$, whereas $\rho(0) = E$ for the case $r \leq D$ ([16]). Thus, the free boundary profile develops an initial jump in the case $r > D > 0$. Notice that the case $0 < r \leq D$ can be also treated by other methods based on integral equations [1, 6, 12, 14]. Kwok [5] derived another integral equation which covers both cases $0 < r \leq D$, as well as $r > D > 0$. However, in the latter case the integral equation becomes singular as $t \rightarrow T^-$, leading to numerical instabilities near expiry.

To transform equation (1) defined on a time dependent spatial domain $(0, S_f(t))$, we introduce the following change of variables [12, 13]:

$$\tau = T - t, \quad x = \ln \left(\frac{\rho(\tau)}{S} \right), \quad \text{where} \quad \rho(\tau) = S_f(T - \tau).$$

Clearly $\tau \in (0, T)$ and $x \in (0, \infty)$ whenever $S \in (0, S_f(t))$. Let us further define the auxiliary function $\Pi = \Pi(x, \tau)$ as follows:

$$\Pi(x, \tau) = V(S, t) - S \frac{\partial V}{\partial S}(S, t). \quad (3)$$

It is shown [6, 12, 14] that under suitable regularity assumptions on the input data the free boundary problem (1), (2) can be transformed into the initial boundary value problem for parabolic PDE:

$$\frac{\partial \Pi}{\partial \tau} + \hat{a}(\tau) \frac{\partial \Pi}{\partial x} - \frac{\sigma^2}{2} \frac{\partial^2 \Pi}{\partial x^2} + r\Pi = 0, \quad x > 0, \quad \tau \in (0, T), \quad (4)$$

$$\Pi(0, \tau) = -E, \quad \Pi(+\infty, \tau) = 0, \quad (5)$$

$$\Pi(x, 0) = \begin{cases} -E, & \text{for } x < \ln \frac{r}{D}, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\hat{a}(\tau) = \frac{\rho'(\tau)}{\rho(\tau)} + r - D - \sigma^2/2$ and

$$\rho(\tau) = \frac{rE}{D} + \frac{\sigma^2}{2D} \frac{\partial \Pi}{\partial x}(0, \tau), \quad \rho(0) = \frac{rE}{D}. \quad (7)$$

We repeat that the problem (1), (2) is a *nonlinear* parabolic equation with a *nonlocal* constraint given by (7).

The solution Π of the problem (4)-(7) is continuous for $t > 0$. The discontinuity appears only at the point $P^* = (\log(r/D), 0)$. The derivatives of the solution exist and are sufficiently smooth in $[0, L] \times [0, T]$, outside of the neighbourhood of P^* .

3 Difference Schemes

In order to solve the problem (4)-(7) numerically, we introduce L which is a large value of x , where we impose the right boundary condition in (5): $\Pi(L, \tau) = 0$. Next, for given positive integers N and M we define the meshes: $\bar{\omega}_h = \{0\} \cup \{L\} \cup \omega_h$, $\omega_h = \{x_i = ih, i = 1, \dots, (N-1), h = L/N\}$ and $\bar{\omega}_k = \{0\} \cup \{T\} \cup \omega_k$, $\omega_k = \{t^j = jk, j = 1, \dots, (M-1), k = T/M\}$. Our goal is to define a finite difference method suitable for computing $y_i^j \approx \Pi(x_i, t_j)$ for $(x_i, t_j) \in \omega_h \times \omega_\tau$ and associated front position $z^j \approx \rho(t_j)$ for $t_j \in \omega_k$. The weighted difference schemes [11] have the following form:

$$\frac{y_i^{j+1} - y_i^j}{k} + \hat{a}(t^{j+1}) \left(\theta_1 \frac{y_{i+1}^{j+1} - y_{i-1}^{j+1}}{2h} + (1 - \theta_1) \frac{y_{i+1}^j - y_{i-1}^j}{2h} \right) \quad (8)$$

$$- \frac{\sigma^2}{2} \left(\theta_2 \frac{y_{i-1}^{j+1} - 2y_i^{j+1} + y_{i+1}^{j+1}}{h^2} + (1 - \theta_2) \frac{y_{i-1}^j - 2y_i^j + y_{i+1}^j}{h^2} \right) = 0,$$

$$y_0^{j+1} = -E, \quad y_N^{j+1} = 0; \quad y_i^0 = \begin{cases} -E, & \text{for } x_i \leq \ln(r/D) \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where

$$\hat{a}(t_{j+1}) = \frac{z^{j+1} - z^j}{kz^{j+1}} + r - D, \quad (10)$$

and

$$z^{j+1} - \left(\frac{rE}{D} + \frac{\sigma^2}{2D} \frac{-3y_0^{j+1} + 4y_1^{j+1} - y_2^{j+1}}{2h} \right) = 0, \quad z^0 = \frac{rE}{D}, \quad (11)$$

or introducing an artificial node in space x_{-1}

$$z^{j+1} = \frac{rE}{D} + \frac{\sigma^2}{2D} \frac{y_1^{j+1} - y_{-1}^{j+1}}{2h}, \quad z^0 = \frac{rE}{D}. \quad (12)$$

Writing the finite difference equations (8) for $i = 1, \dots, N-1$ and introducing the boundary conditions from (9) and the discretization of the moving boundary (11) or (12), an algebraic nonlinear system of equations results. In [9] the authors

apply implicit finite difference scheme, semi-implicit scheme and upwind explicit scheme for the American put option, combining with the penalty method. The time step parameter for the explicit case is much smaller ($k = 5.0 \cdot 10^{-6}$). Therefore in this work we consider the case of fully implicit scheme, i.e. $\theta_1 = \theta_2 = 1$.

4 Iterative Algorithms

In order to solve the nonlinear system of algebraic equations we developed the following algorithms.

Algorithm 1. This algorithm is based on the *regula-falsi method* and consists in the following steps.

Step 1. Let the solution on the time level t_j be known. Let also $\Delta = z_{j+1} - z_j$. For a fixed time step k and suitable chosen initial values Δ_1^l and Δ_2^l we find $\hat{a}(t^{j+1})$ from (10).

Step 2. Then we solve the linear system (8), (9), (12) without Dirichlet boundary condition $y_0^{j+1} = -E$ from (9) for both values of Δ_s^l , $s = 1, 2$. The corresponding solutions we denote by $y_i^{j+1}(\Delta_s^l)$.

Step 3. We want the obtained in Step 2 solution to satisfy the Dirichlet boundary condition $y_0^{j+1} = -E$. So, we check if for $y_0^{j+1}(\Delta_s^l)$, $s = 1, 2$ the conditions are fulfilled:

$$\|y_0^{j+1}(\Delta_s^l) + E\| < tol.$$

If not, we find the new value Δ_3^l by the formula

$$\Delta_3^l = \frac{y_0^{j+1}(\Delta_2^l)(z_j + \Delta_1^l) - y_0^{j+1}(\Delta_1^l)(z_j + \Delta_2^l)}{y_0^{j+1}(\Delta_2^l) - y_0^{j+1}(\Delta_1^l)} - z_j.$$

Step 4. Discard the value Δ_s^l , that corresponds to the largest of the two values $|y_0^{j+1}(\Delta_s^l) + E|$, $s = 1, 2$. With the remaining value Δ_s^l and including Δ_3^l as initial values on the $l + 1$ -th iteration Δ_1^{l+1} , Δ_2^{l+1} we repeat the Step 1.

Algorithm 2. We now describe an algorithm based on the *Newton method*.

Step 1. We eliminate the known boundary values $y_0^{j+1} = -E$ and $y_N^{j+1} = 0$ in (8) and adding (11) we obtain a nonlinear system for N unknowns: y_i^{j+1} , $i = 1, 2, \dots, N - 1$ and z_{j+1} . We denote by $\mathbf{Y}^l$ the vector of this unknowns on the l -th iteration.

Step 2. We use Newton method in the following form:

$$\mathbf{J}^l (\mathbf{Y}^{l+1} - \mathbf{Y}^l) = -\mathbf{F}^l, \quad (13)$$

where the Jacobian matrix is $\mathbf{J} = \begin{pmatrix} \overset{l}{\mathbf{J}}_{11} & \overset{l}{\mathbf{J}}_{12} \\ \overset{l}{\mathbf{J}}_{21} & \overset{l}{\mathbf{J}}_{22} \end{pmatrix}$

$$\overset{l}{\mathbf{J}}_{11} = \begin{pmatrix} c & b & & & \\ a & c & b & & \\ & \ddots & \ddots & \ddots & \\ & & a & c & b \\ & & & a & c \end{pmatrix} \quad \overset{l}{\mathbf{J}}_{12} = \begin{pmatrix} \frac{\partial a}{\partial z_{j+1}}(-E) + \frac{\partial b}{\partial z_{j+1}}y_2^{j+1} \\ \frac{\partial a}{\partial z_{j+1}}y_1^{j+1} + \frac{\partial b}{\partial z_{j+1}}y_3^{j+1} \\ \vdots \\ \frac{\partial a}{\partial z_{j+1}}y_{N-3}^{j+1} + \frac{\partial b}{\partial z_{j+1}}y_{N-1}^{j+1} \\ \frac{\partial a}{\partial z_{j+1}}y_{N-2}^{j+1} \end{pmatrix} \quad \overset{l}{\mathbf{J}}_{21} = \begin{pmatrix} \frac{-\sigma^2}{Dh} \\ \frac{\sigma}{4Dh} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}^T$$

and $\overset{l}{\mathbf{J}}_{22} = 1$. Similarly $\overset{l}{\mathbf{Y}} = \begin{pmatrix} \overset{l}{\mathbf{Y}}_{11} & \overset{l}{\mathbf{Y}}_{12} \end{pmatrix}^T$, $\overset{l}{\mathbf{Y}}_{11} = (y_1^{j+1}, \dots, y_{N-1}^{j+1})$, $\overset{l}{\mathbf{Y}}_{12} = z_{j+1}$.

The elements of the matrix $\overset{l}{\mathbf{J}}_{11}$ are

$$a = -\frac{1}{2h} \left(\frac{z_{j+1} - z_j}{kz_{j+1}} + r - D - \frac{\sigma^2}{2} \right) - \frac{\sigma^2}{2h^2}, \quad c = \frac{1}{k} + \frac{\sigma^2}{h^2} + r,$$

$$b = \frac{1}{2h} \left(\frac{z_{j+1} - z_j}{kz_{j+1}} + r - D - \frac{\sigma^2}{2} \right) - \frac{\sigma^2}{2h^2}.$$

This iteration process is done until $\max(|\overset{l+1}{\mathbf{Y}} - \overset{l}{\mathbf{Y}}|) < tol$. The matrix $\overset{l}{\mathbf{F}}$ is obtained from (8) and (12) after substitution $\overset{l}{\mathbf{Y}}$ in the left hand side.

Step 3. The solution on the $(j+1)$ -th time layer is taken as initial iteration for the next time layer.

For solving (13) we do the following stages. First, we consider the matrix equation

$$\overset{l}{\mathbf{J}}_{11} \overset{l+1}{\mathbf{Y}}_{11} = -\overset{l}{\mathbf{F}}_{11} + \overset{l}{\mathbf{J}}_{11} \overset{l}{\mathbf{Y}}_{11}.$$

The matrix $\overset{l}{\mathbf{J}}_{11}$ is a threedagonal matrix and we apply the Thomas algorithm to find $\overset{l+1}{\mathbf{Y}}_{11}$. Second, we solve

$$\overset{l}{\mathbf{J}}_{12} \overset{l+1}{\mathbf{Y}}_{11} + \overset{l}{\mathbf{J}}_{22} \overset{l+1}{\mathbf{Y}}_{12} = -\overset{l}{\mathbf{F}}_{12}.$$

5 Numerical Experiments

Example 1. We consider problem (1) with parameter values $E = 10$, $r = 0.1$, $D = 0.05$ and $T = 1$, see [6, 12–15].

As there is not an analytical solution to the proposed free boundary problem, we chose as an exact solution the numerical solution with a small mesh parameter

Table 1. Mesh-refinement analysis of the regula-falsi method for Example 1

N	M	E_{∞}^N	m	ρ_{end}	E_{ρ}^N	m	l_{max}
50	0.001	0.1335	-	22.2979	0.2131		2
100	0.001	0.0479	1.4787	22.3563	0.1156	0.8824	2
200	0.001	0.0268	0.8378	22.3707	0.0631	0.8734	2
400	0.001	0.0216	0.3112	22.3743	0.0312	1.0161	2
800	0.001	0.0203	0.0896	22.3752	0.0203	0.6201	2

Table 2. Mesh-refinement analysis of the Newton method for Example 1

N	M	E_{∞}^N	m	ρ_{end}	E_{ρ}^N	m	l_{max}
50	0.001	1.1061	-	22.2978	0.2126	-	3
100	0.001	0.5753	0.9431	22.3559	0.1036	1.0371	3
200	0.001	0.2859	1.0088	22.3706	0.0467	1.1495	4
400	0.001	0.1350	1.0826	22.3744	0.0182	1.3595	4
800	0.001	0.0582	1.2139	22.3753	0.0058	1.6498	5

$h = 1/3200$ (i.e. $N=3200$). We denote the error of the computed solution in maximum norm by

$$E_{\infty}^N = \max |y_{i_c, N}^M - y_{i_c, 3200}^M|,$$

where i_c is the number of the nodes, common for the both meshes with $h = 1/N$ and $h = 1/3200$. In Table 1 we give mesh-refinement analysis for the numerical solution, obtained with the regula-falsi method. We control the error of the solution and the error of the free boundary $\rho(t)$:

$$E_{\rho}^N = \max |z_N^M - z_{3200}^M|.$$

Also, the rate of convergence m , $m = \log(E_{\infty}^N/E_{\infty}^{2N})/\log 2$, and the number of maximum iterations l_{max} are presented. For the Newton method the results show first order of accuracy for the solution and for the moving boundary it increases with respect of N . For the regula-falsi method the rate of convergence decreases for the solution and for the moving boundary it is near 1. But the absolute values of the errors are smaller in the second method. The final values of the free boundary ρ in both methods are close to the dose, obtained for the same problem in [6, 12]. In Fig. 1a) the profile of the free boundary ρ is presented. In Fig. 1b) the numerical solution of the portfolio Π obtained by the Newton method is depicted.

In Fig. 2a) the free boundary position for a long time $T = 50$ years is shown. The obtained values are: for $N = 50$ $\rho = 36.7628$; $N = 200$ $\rho = 36.8122$; $N = 800$ $\rho = 36.8156$. Another interesting case is when the dividend D is close to the rate r . In Fig. 2b) the free boundary position for $T = 0.01$, $\tau = 0.00001$, $D = 0.09$ and $r = 0.1$ is presented. For $N = 50$ $\rho = 11.2429$; $N = 800$ $\rho = 11.2536$; $N = 3200$ $\rho = 11.2537$.

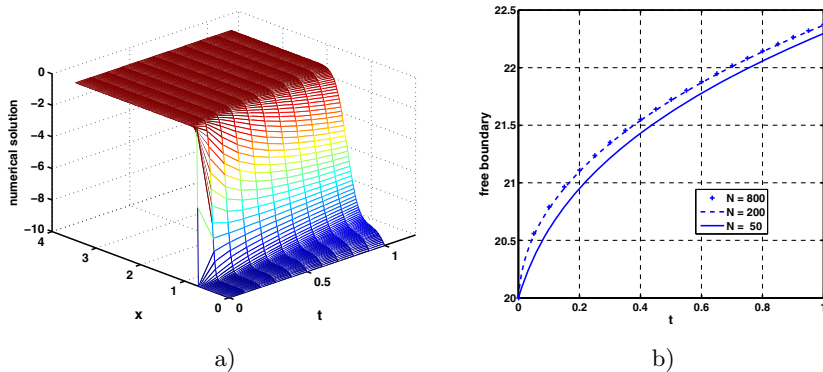


Fig. 1. (a) The numerical solution for $N = 200$, $M = 1000$; (b) The free boundary position for $N = 50$, $N = 200$, $N = 800$

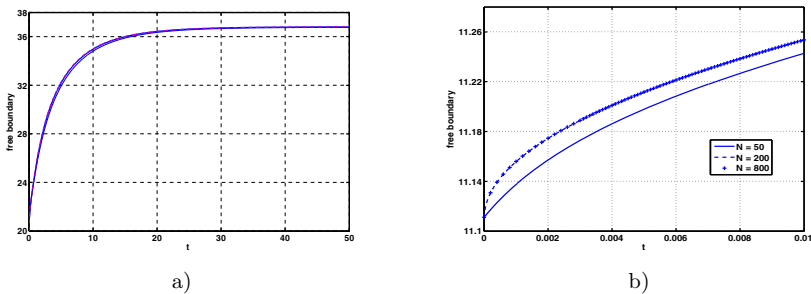


Fig. 2. (a) The free boundary position for $t = 50$ years; (b) The free boundary with dividend close to rate

6 Conclusions

In this communication on the base of a weight difference scheme we have developed two algorithms for solving a free boundary value problem, known in the literature as the Black-Scholes equation for pricing the American call options. To solve this degenerate parabolic problem we use Landau's transformation which fixes the moving interface. The two algorithms choose a constant time step. The first one calculates the sequence of parameters until the moving boundary condition is satisfied. The second algorithm uses Newton's method. Advantage of the both algorithms is the convergence with only a few iterations. Nevertheless the approximations are of second order, due to the discontinuity of the initial data the results show near first order of the methods. More careful analysis and smoothing techniques like Rannacher [10] procedure are objective in a future work.

Acknowledgements. We thank Prof. D. Sevcovic for the useful discussion of the paper. This research is supported by the Bulgarian National Fund of Science under Project Bg-Sk-203/2008.

References

1. Bokes, T., Sevcovic, D.: Early exercise boundary for American type of floating strike Asian option and its numerical approximation (2009) (submitted)
2. Broadie, M., Dempse, J.: American option valuation: New bounds, approximations and comparison of existing methods. *Review of Financial Studies* (1994)
3. Gupta, S.C.: *The Classical Stefan Problem: Basic Concepts, Modelling and Analysis*. North-Holland Series in Applied Mathematics and Mechanics. Elsevier, Amsterdam (2003)
4. Han, H., Wu, X.: A Fast Numerical Method for the Black-Scholes Equation of American Options. *SIAM J. Numer. Anal.* 41(6), 2081–2095 (2003)
5. Kwok, J.K.: *Mathematical Models of Financial Derivatives*. Springer, Heidelberg (1998)
6. Lauko, M., Sevcovic, D.: Comparison of Numerical and Analytical Approximations of the Early Exercise Boundary of the American Put Option (2010) (submitted)
7. Meirmanov, A.M.: *The Stefan Problem*. Walter de Gruyter, Berlin (1992)
8. Moyano, E., Scarpentini, A.: Numerical Stability Study and Error Estimation for Two Implicit Schemes in a Moving Boundary Problem. *Num. Meth. Part. Diff. Eq.* 16(1), 42–61 (2000)
9. Nielsen, B., Skavhaug, O., Tveito, A.: Penalty and Front-fixing Methods for the Numerical Solution of American Option Problems. *J. of Comp. Fin.* 5(4), 69–97 (2002)
10. Rannacher, R.: Discretization of the Heat Equation with Singular Initial Data. *Zeit. Ang. Math. Methods (ZAMM)* 62, 346–348 (1982)
11. Samarskii, A.A.: *The Theory of Difference Schemes*. Marcel Dekker, New York (2001)
12. Sevcovic, D.: Analysis of the Free Boundary for the Pricing of an American Call Option. *Eur. J. Appl. Math.* 12, 25–37 (2001)
13. Sevcovic, D.: Transformation Methods for Evaluating Approximations to the Optimal Exercise Boundary for Linear and Nonlinear Black-Scholes Equations. In: Ehrhard, M. (ed.) *Nonlinear Models in Mathematical Finance: New Research Trends in Optimal Pricing*, pp. 153–198. Nova Sci. Publ., New York (2008)
14. Stamicar, R., Sevcovic, D., Chadam, J.: The Early Exercise Boundary for the American Put Near Expiry: Numerical Approximation. *Canadian Applied Mathematics Quarterly* 7(4), 427–444 (1999)
15. Tangman, D.Y., Gopaul, A., Bhuruth, M.: A Fast High-order Finite Difference Algorithms for Pricing American Options. *J. Comp. Appl. Math.* 222, 17–29 (2008)
16. Wilmott, P., Dewynne, J., Howison, S.: *Option Pricing, Mathematical Models and Computation*. Oxford Financial Press (1993)
17. Zhu, Y., Ren, H., Xu, H.: Improved Effectiveness Evaluating American Options by the Singularity-separating Method. Techn. report, Univ. of North Carolina at Charlotte (1997)

A Numerical Study of a Parabolic Monge-Ampère Equation in Mathematical Finance

Miglena N. Koleva and Lubin G. Vulkov

Faculty of Natural Science and Education
University of Rousse, 8 Studentska str., Rousse 7017, Bulgaria
`{mkoleva,lvalkov}@uni-ruse.bg`

Abstract. We propose iterative algorithms for solving finite difference schemes approximating an initial value problem of a parabolic Monge-Ampère equation, arising from the optimal investment of mathematical finance theory. We investigate positivity and convexity preserving properties of the numerical solution. Convergence results are also given. Numerical experiments demonstrate the efficiency of the algorithms and verify theoretical statements.

1 Introduction

The following initial value problem was derived in [1,5,9]

$$\begin{aligned} V_t V_{xx} + r x V_x V_{xx} - \lambda V_x^2 &= 0, \quad (x, t) \in \mathbf{R} \times [0, T], \\ V(x, T) &= g(x), \quad x \in \mathbf{R}, \quad V_{xx} < 0, \quad g'(x) > 0. \end{aligned} \quad (1)$$

$V = V(x, t)$ is called a *value function* (depending on underlying asset x and the time t) of the market model presented in [9, Chapter 2], V_{xx} is the Gamma of the option. The coefficient $\lambda = (c - r)/\varrho > 0$ where the positive constants r , c and ϱ are the *interest rate*, the *appreciation rate* and the *volatility* ($c - r > 0$), respectively. The model (1) describes the *simple market model in the case of one asset*. In a typical case, function $g(x)$ is given by

$$g(x) = 1 - e^{-\mu x}, \quad \mu > 0. \quad (2)$$

Often in the mathematical finance models, the "terminal condition" is changed into the "initial condition" by setting

$$V(x, T - t) = v(x, t). \quad (3)$$

Now, by substitution [8]

$$u(x, t) = -\frac{v_x(x, t)}{v_{xx}(x, t)}, \quad (x, t) \in \mathbf{R} \times [0, T], \quad (4)$$

we get the following initial-value (Cauchy) problem

$$\begin{aligned} u_t &= \lambda u^2 u_{xx} + r x u_x - r u, \quad (x, t) \in \mathbf{R} \times (0, T], \\ u|_{t=0} &= -\frac{g'}{g''} = u_0, \quad x \in \mathbf{R}, \end{aligned} \quad (5)$$

where $u_0 > 0$. Looking at the angle of economic mathematics we need the Lipschitz continuity of u_0 and from pure mathematics point of view, this is a necessary condition for the existence of solutions, see [8]. In this work (Theorem 2, page 62), under some additional assumptions for u_0 , namely $u_0 \in C^\gamma(\mathbf{R})$, for some number $0 < \gamma < 1$ and

$$C_1(1+x^2)^\beta \leq u_0(x) \leq C_2(1+x^2)^{1/2}, \quad (6)$$

the author proves the existence of *classical solutions* to the initial value problem (5), satisfying the next inequality for $(x, t) \in \mathbf{R} \times [0, T]$

$$\underline{v}(x, t; C_1, C_2) = C_1(d+x^2)^\beta e^{-M_1 t} \leq u(x, t) \leq \overline{v}(x, t; C_2) = C_2(e^{M_2 t} + x^2)^{\frac{1}{2}}, \quad (7)$$

where $C_1 > 0$, $C_2 > 0$, $-\infty < \beta \leq \frac{1}{2}$, $M_1 = 4(|\beta| + 1)^2 C_2^2 + (2|\beta| + 1)|r|$, $M_2 = 2C_2^2 + 2|r|$ and $d = e^{M_2 t}$.

As we will use (in the proof of Theorem 1) the proof of Theorem 2.1 [8], we outline its three steps. Denote $B_R = (-R, R)$, $Q_{R,T} = B_R \times (0, T)$ and let $u \in C(\overline{Q}_T)$, $m \leq u \leq M$, m, M are constants. Consider the following IBVP:

$$\begin{aligned} -v_t + u^2 v_{xx} + r x v_x - r v &= 0, \quad (x, t) \in Q_{R,T}, \\ v(\pm R, t) &= u_0(\pm R), \quad t \in [0, T], \\ v(x, 0) &= u_0(x), \quad x \in B_R. \end{aligned} \quad (8)$$

By the Hölder estimates for nondivergence form equations ([4], Theorem 7, page 137 and Theorem 5, page 165), for each solution of (8) v , there exist constants $\alpha_0, C_3 > 0$ such that $\|v\|_{C^{\alpha_0, \alpha_0/2}(\overline{Q}_{R,T})} \leq C_3$, where α_0 and C_3 are constants, depending only on $m, M, \|u_0(x)\|_{C^\gamma}$ and T . Let $\alpha = \alpha_0/2$ and

$$K_R = \{u \in C^{2+\alpha, 1+\alpha/2}(Q_{R,T}) \cap C^{\alpha, \alpha/2}(\overline{Q}_{R,T}), \quad \underline{v}(x, t; C_1) \leq u(x, t) \leq \overline{v}(x, t; C_2)\}.$$

Step 1. For each $u \in K_R$, by the theory of linear equations, there exists a unique solution v , for which $\underline{v}, \overline{v}$ are lower and upper solution, respectively and the following estimate holds

$$\|v\|_{C^{2+\alpha, 1+\alpha/2}(Q_{R,T})} \leq C_4, \quad C_4 = \text{const.}$$

Step 2. The mapping $\Phi : K_R \rightarrow K_R$, where $\Phi(u) = v$ is the solution of IBVP (8) corresponding to u as its coefficient in the differential equation. The mapping Φ has a fixed point, that there is a function $u \in K_R$, satisfying the IBVP

$$\begin{aligned} -u_t + \lambda u^2 u_{xx} + r x u_x - r u &= 0, \quad (x, t) \in Q_{R,T}, \\ u(\pm R) &= u_0(\pm R), \quad t \in [0, T], \\ u(x, 0) &= u_0(x), \quad x \in B_R = (-R, R) \times (0, T), \quad Q_{R,T} = B_R \times (0, T). \end{aligned} \quad (9)$$

Step 3. There is a solution $u(x, t)$ to (5) satisfying $\underline{v} \leq u \leq \overline{v}$. The solution u is the limit of nested sequence $\{u_n\}$, where $u_n \in K_n$ is the solution given in *step 2* to IBVP (8). It converges pointwise as well as its partial derivatives (first-order time derivative and second order space derivative).

Further, we will concentrate on the numerical investigation of problem (5). The rest of the paper is organized as follows. In Section 2 we study the monotone convergence of the space approximation of problem (5). Also, some properties of the semidiscrete solution are established. In Section 3 the time approximation is discussed. Iterative methods for solution of the nonlinear difference equations are developed in Section 4. Numerical experiments are presented in Section 5.

2 Finite Difference Schemes

We consider the problem (5) in the finite interval $B(R) = [-R, R] \subset \mathbf{R}$ with boundary conditions $u(-R, t) = u_0(-R)$ and $u(R, t) = u_0(R)$. The domain $B(R)$ is discretized by uniform mesh $\bar{\omega}_h$,

$$\bar{\omega}_h = \{x_i \mid x_i = -R + (i-1)h, \quad h = 2R/(N-1), \quad i = 1, \dots, N\}.$$

Denote the numerical solution at point (x_i, t) by $y_i = y(x_i, t)$ and central differences $y_{\bar{x},i} = (y_{i+1} - 2y_i + y_{i-1})/h^2$, $y_{\hat{x},i} = (y_{i+1} - y_{i-1})/(2h)$. The boundary value problem (9) is approximated by the difference scheme

$$\begin{aligned} y_{t,i} &= \lambda y_i^2 y_{\bar{x},i} + r x_i y_{\hat{x},i} - r y_i, \quad i = 2, \dots, N-1, \\ y_1 &= u_0(-R), \quad y_N = u_0(R), \\ y_i(0) &= u_0(x_i), \quad i = 1, \dots, N. \end{aligned} \tag{10}$$

Theorem 1. *Assume that the initial function $u_0(x) \in C^\gamma(R)$ for some $0 < \gamma < 1$ and the inequality (7) is fulfilled for $t = 0$. Then, there exists unique smooth solution $\{y_i(t)\}_{i=1}^N$ to the problem (10). Also, the linear interpolant $y^I(x, t)$ of this solution converges to $u_R(x, t)$ as $N \rightarrow \infty$ and the next estimate holds*

$$\|y^I - u\|_C \leq Ch^2,$$

where C is a constant independent of h .

Proof (Outline). We follows the strategy of step 1- step 3. For each $u \in K_R$ we define the auxiliary semidiscrete problem

$$\begin{aligned} z_{t,i} &= \lambda u^2(x_i, t) z_{\bar{x},i} + r x_i z_{\hat{x},i} - r z_i, \quad i = 2, \dots, N-1, \\ z_1 &= u_0(-R), \quad z_N = u_0(R), \\ z_i(0) &= u_0(x_i), \quad i = 1, \dots, N, \end{aligned} \tag{11}$$

where $u(x, t)$ is the solution of problem (9). By the theory of difference schemes for linear parabolic equations [7], $\|z - v\| \leq Ch^2$. By the theory of ODEs there exists unique solution $z \in C^{1+\alpha/2}(0, T]$ of IVP (10). Define the ball

$$\begin{aligned} K_R^N &= \{y^N(t) = (y_1(t), \dots, y_N(t)), \quad y_i(t) \in C^{1+\alpha/2}[0, T], \\ &\quad \underline{y}(x_i, t) \leq y_i(t) \leq \bar{v}(x_i, t), \quad t \in [0, T], \quad i = 1, \dots, N\}. \end{aligned}$$

Define the map $\Phi^N : K_R^N \rightarrow K_R^N$, where $\Phi^N(u) = z$ is the solution the IVP (11). The mapping Φ^N has a fixed point, that is, there is a function $y^N(t)$ satisfying the IVP (10). $\square$

Now, we derive the full discretization of problem (5) in $[-R, R] \times [0, T]$ by uniform mesh $\overline{\omega} = \overline{\omega}_h \times \overline{\omega}_\tau$, $\overline{\omega}_\tau = \{t^n \mid t^n = n\tau, \ n = 0, 1, \dots\}$.

Denoting the numerical solution at point (x_i, t^n) by $y_i^n = y(x_i, t^n)$ and $y_{i,i}^n = (y_i^n - y_i^{n-1})/\tau$, we obtain from (10) the following weighted ($\theta = \{0, 1\}$) discretization of problem (5) for $n = 0, 1, \dots$

$$\begin{aligned} y_{t,i}^{n+1} &= \lambda[\theta(y_i^{n+1})^2 + (1-\theta)(y_i^n)^2]y_{xx,i}^{n+1} + rx_i y_{x,i}^{n+1} - ry_i^{n+1}, \quad i = 2, \dots, N-1, \\ y_1^{n+1} &= u_0(-R), \quad y_N^{n+1} = u_0(R), \\ y_i^0 &= u_0(x_i), \quad i = 1, \dots, N. \end{aligned} \quad (12)$$

In the next theorem we establish some important (from mathematical finance point of view) properties of the numerical solution: *positivity and convexity preservation* [2,3].

Theorem 2. *Let $u_0(x)$ be a positive initial function chosen as in (6) and*

$$h < \frac{\lambda}{r} \min_{1 \leq i \leq N} \frac{(y_i^n)^2}{x_i}. \quad (13)$$

Then the finite difference discretization (12) results in a positive solution on each time level. Also, the numerical solution v_i^n of (3) is a convex function at each time level.

Proof (outline). The positivity of y^n follows from the discrete maximum principle. On the base of Theorem 1, we can rewrite the scheme (12) in the form $y_{t,i}^{n+1} = \lambda[C\theta(\tau + h^2) + (y_i^n)^2]y_{xx,i}^{n+1} + rx_i y_{x,i}^{n+1} - ry_i^{n+1}$, $i = 2, \dots, N-1$, where $C\theta(\tau + h^2) + (y_i^n)^2 > (y_i^n)^2/2 > u_0^2(x_i)/2$. Then by induction with respect to $n = 0, 1, \dots$, we show that $\min\{u_0(-R), u_0(R)\} \leq y_i^n \leq \max\{u_0(-R), u_0(R)\}$. Next, from (4), at some time level t^n , $n = 1, 2, \dots$, we have

$$v_x = e^{-\int_a^x u(\rho, t) d\rho} \quad \text{or} \quad v_{x,i}^n = e^{-(\frac{h}{2}y_1^n + h \sum_{j=1}^{i-1} y_j^n + \frac{h}{2}y_i^n)} > 0.$$

Thus, because of $v_{xx,i}^n = -y_i^n v_{x,i}^n$ and positivity of the solution y_i^n we conclude that $v_{xx,i}^n < 0$. $\square$

3 Iterative Processes

In the computations one can follow the iterative scheme *step 1 - step 3*, but the convergence is very slow. On the base of (12), $\theta = 1$ we organize an iteration process for $k = 0, 1, \dots$, at each time level t^n , $n = 0, 1, \dots$

Gauss-Seidel type

$$\begin{aligned} y_i^{(k+1)} &= \tau \lambda (y_i^{(k)})^2 y_{xx,i}^{(k+1)} + \tau r x_i y_{x,i}^{(k+1)} - \tau r y_i^{(k+1)} + y_i^n, \quad i = 2, \dots, N-1, \\ y_1^{(k+1)} &= u_0(-R), \quad y_N^{(k+1)} = u_0(R), \\ y_i^{(0)} &= y_i^n, \quad i = 1, \dots, N, \quad n = 0, 1, \dots \end{aligned} \quad (14)$$

The iterations continued until we reach some tolerance between the solution of two subsequent iterations, then $y^{n+1} := y^{(k+1)}$.

Theorem 3. *Suppose that the difference scheme (12), $\theta = 1$ is already solved for $Y^n = [y_1^n, \dots, y_N^n]^T$ in the n -th time layer. Then, for sufficiently small τ and h , satisfied (13), there exists unique solution Y^{n+1} of (12), $\theta = 1$ and the iterative process is convergent with first order rate of convergence.*

Proof (Outline). First, writing (12), $\theta = 1$ as a nonlinear operator equation [6] we prove existence of unique solution Y^{n+1} . Then, subtracting the first equation of (12), $\theta = 1$ from (14) we get

$$\Delta y_i^{(k+1)} - \tau \lambda (y_i^{(k)})^2 \Delta y_{\bar{x},i}^{(k+1)} - \tau r x_i \Delta y_{\bar{x},i}^{(k+1)} + \tau r \Delta y_i^{(k+1)} = \tau \lambda y_{\bar{x},i}^{n+1} (y_i^{(k)} + y_i^{n+1}) \Delta y_i^{(k)}.$$

Applying the discrete maximum principle we find the estimate

$$\|\Delta y^{(k+1)}\|_C \leq q \|\Delta y^{(k)}\|_C, \quad q < 1, \quad \text{independent of } \tau \text{ and } h. \quad \square$$

Newton's method

At each time level, for $\sigma \in (0, 1]$ we seek the correction $\delta y^{(k+1)}$, where $y^{(k+1)} = y^{(k)} + \delta y^{(k+1)}$, from the equation

$$\begin{aligned} a_i^{(k)} \delta y_{i-1}^{(k+1)} + \tilde{b}_i^{(k)} \delta y_i^{(k+1)} + c_i^{(k)} \delta y_{i+1}^{(k+1)} &= \sigma d_i^{(k)} + (1 - \sigma) d_i^n - y_{t,i}^{(k)}, \quad i = 2 \dots N-1, \\ \delta y_1^{(k+1)} &= 0, \quad \delta y_N^{(k+1)} = 0, \\ y_i^{(0)} &= y_i^n, \quad i = 1, \dots, N, \quad n = 0, 1, \dots \end{aligned} \quad (15)$$

$$\text{where } a_i^{(k)} = \sigma \left[-\frac{\lambda}{h^2} (y_i^{(k)})^2 + \frac{r x_i}{2h} \right], \quad b_i^{(k)} = \sigma \left[\frac{1}{\tau} + \frac{2\lambda}{h^2} (y_i^{(k)})^2 + r \right], \quad \tilde{b}_i^{(k)} = b_i^{(k)} - 2\lambda y_i^{(k)} y_{\bar{x},i}^{(k)}, \quad c_i^{(k)} = \sigma \left[-\frac{\lambda}{h^2} (y_i^{(k)})^2 - \frac{r x_i}{2h} \right] \text{ and } d_i^{(k)} = \lambda (y_i^{(k)})^2 y_{\bar{x},i}^{(k)} + r x_i y_{\bar{x},i}^{(k)} - r y_i^{(k)}.$$

Theorem 4. *Let the conditions of Theorem 2 are fulfilled and C_1 and C_2 are constants, such that $\|\varphi'\|_C \geq C_1 > 0$, $\|\varphi''\|_C \leq C_2$, where $\varphi(y) = -1/y$. Assume that $\|y^{(0)} - y^{n+1}\|_C \leq \rho$ and $2\rho C_1/C_2 < 1$. Then for the solution of (15), $\sigma = 1$ we have*

$$\|\Delta y^{(k+1)}\|_C \leq 2C_1/C_2 (2C_1 C_2 \rho)^{2^k}, \quad k = 0, 1, \dots$$

Proof (Outline). Rewriting the equation (5) in the form

$$\frac{\partial \varphi(u)}{\partial t} - \lambda \frac{\partial^2 u}{\partial x^2} - r x \frac{\partial \varphi(u)}{\partial x} - r \varphi(u) = 0, \quad (16)$$

The full discretization of (16) takes the form

$$\varphi(y_i^{n+1}) - \tau \lambda y_{\bar{x},i}^{n+1} \frac{\varphi(y_{i+1}^{n+1}) - \varphi(y_{i-1}^{n+1})}{2h} - \tau r x_i \varphi(y_i^{n+1}) = \varphi(y_i^n).$$

Further, we apply a quasilinearization combined with discrete maximum principle and induction. $\square$

Similarly, the same results can be proved in the case of $\sigma = 0.5$. From $\varphi''(y) < 0$ and the discrete maximum principle follows that $y^{(k+1)} \leq y^{n+1}$ and therefore the numerical solution (15) approximates the exact solution of (5) *from below*.

Remark 1. For iteration process of Newton and Gauss-Seidel type the statement of Theorem 2 is also true and the proof is based on the discrete maximum principle.

4 Numerical Experiments

In this section we will verify the convergence rate (in maximal and L_2 discrete norms) of numerical schemes (12), (14), (15). The iterations for all methods continued until the maximal solution's difference between two subsequent iterations is less than 10^{-12} . The computations are performed in the interval $[-3, 3]$ for $\lambda = 1$, $r = 0.5$, $T = 1$. Numerical experiments show that for positivity preservation property of the numerical solution, obtained by iteration methods there is no restriction (13).

Example 1. (*Exact solution*) We add a right hand side f in (5), (after that in (12)-(15)) and determine input datum, such that the positive in $[-3, 3]$ function $u = e^{-t}(-x^2 + x + 12)$ is the exact solution of the problem (5). Thus the convergence rate is calculated using double mesh principle

$$CR = \log_2(E^N / E^{2N}), \text{ where } E^N = \max_i |u_i^n - y(x_i, t^n)|, i = 1, \dots, N.$$

To show the convergence rate of the numerical solution, we chose $\tau = h^2$ for the iteration algorithms (12),(14) and (15) for $\sigma = 1$; $\tau = h$ for (15) with $\sigma = 0.5$. The results - errors, convergence rate and CPU times are listed in Table 1. As

Table 1. Errors (E^N), convergence rates (CR) and CPU times, Example 1

Scheme (12), $\theta = 0$			Scheme (14)		Scheme (15), $\sigma = 1$		Scheme (15), $\sigma = 0.5$	
N	E^N	CPU	E^N	CPU	E^N	CPU	E^N	CPU
21	2.2386e-1	0.156	1.9058e-2	0.610	1.9058e-2	0.312	3.5335e-3	0.156
41	5.7264e-2	0.344	4.6635e-3	2.140	4.6635e-3	1.813	8.0200e-4	0.484
CR	1.9669		2.0903		2.0903		2.1394	
81	1.4211e-2	1.422	1.1663e-3	13.540	1.663e-3	11.859	1.9029e-4	1.953
CR	2.0107		1.9995		1.9995		2.0754	
161	3.5631e-3	7.609	2.9052e-4	36.921	2.9052e-4	76.016	4.8843e-5	10.734
CR	1.9958		2.0052		2.0052		1.9620	
321	8.9123e-4	57.015	7.2597e-5	221.481	7.2597e-5	490.125	1.2044e-5	58.266
CR	1.9993		2.0007		2.0007		2.0198	

can be expected, the implicit-explicit scheme (12) is less accurate. The accuracy of both methods: (14) and full implicit scheme (15), $\sigma = 1$ is one and the same results, but the first one is more effective (in the sense of computational time) for fine meshes. The convergence rate of these first three schemes is $\mathcal{O}(\tau + h^2)$. The best results are obtained by full implicit scheme (15), $\sigma = 0.5$, which is not a surprise, as it is $\mathcal{O}(\tau^2 + h^2)$ method.

Example 2. (*Problem (5)*) Now, we test the convergence rate in maximal and L_2 discrete norms of the original problem (5) with iteration procedure (14). Taking into account (2), we chose the initial function $u_0(x) \equiv \mu = 2$. The convergence rate is calculated on three consequently meshes: if y_N^n is the solution at n -th time layer, computed on the space mesh with N grid nodes

$$CR = \log_2(\|y_{N/2}^n - y_N^n\|_h / \|y_N^n - y_{2N}^n\|_h), \text{ where } \|\cdot\|_h \text{ is a maximal or } L_2 \text{ norm.}$$

The ratio $\tau/h^2 = 1$ is fixed. The results are given in Table 2.

Table 2. Convergence rate in max and L_2 discrete norms, Example 2

N	h	max norm	L_2 norm
21	0.3	1.8810	1.8801
41	0.15	1.9866	1.9903
81	0.075	1.9959	1.9969
161	0.0375	1.9988	1.9991
321	0.0187	1.9979	1.9979

Example 3. (*Problem (3)*) From (4), if the solution u at some time level t^n is known, we can find the solution of problem (3) at the same time stage

$$v(x, t^n) = \int_a^x e^{-\int_a^\rho u(s, t^n) ds} d\rho. \quad (17)$$

Taking into account, that the we have the numerical solution $y := y^n$ of u , we need a discrete analogue of (17). For large computational interval it is more suitable to discretize (4) and use (17) to obtain boundary conditions

$$y_i v_{\bar{x}, i} + v_{\bar{x}, i} = 0, \quad i = 2, \dots, N-1, \quad t = t^n, \\ v(a, t^n) = 0,$$

$$v(b, t^n) = \frac{h}{2} e^{-\frac{h}{2} y_1} + h \sum_{j=2}^{N-1} e^{-\left(\frac{h}{2} y_1 + h \sum_{\substack{k=2 \\ k>2}}^{j-1} y_k + \frac{h}{2} y_j\right)} + \frac{h}{2} e^{-\left(\frac{h}{2} y_1 + h \sum_{k=2}^{N-1} y_k + \frac{h}{2} y_N\right)}.$$

On Figure 1 y , v , $y_{\bar{x}}$ and $v_{\bar{x}}$ for $N = 81$, $\tau = 0.05$ are plotted. The initial condition u_0 is the same as in Example 2.

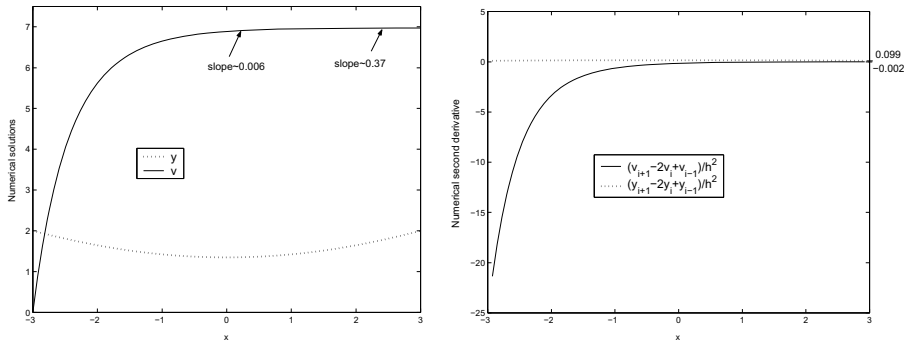


Fig. 1. Numerical solution y and v and second derivatives at $T = 1$, Example 3

5 Conclusions

In this work we presented a second order finite difference methods: implicit-explicit method (12), iteration algorithms (14) and (15) for solving initial value problem of the parabolic Monge-Ampère equation, arising from the optimal investment of mathematical finance theory. We emphasize the advantages of the iteration algorithms: computational efficiency, algorithmically simple for implementation and easy for theoretical investigation. We show that the numerical methods preserve convexity of the solution of the original problem.

Acknowledgement. This research is supported by the Bulgarian National Fund of Science under Projects Sk-Bg-203 and ID_09_0186.

References

1. Bakstein, D., Howison, S.: An arbitrage-free liquidity model with observable parameters for derivatives. In: Working paper. Math. Inst., Oxford Univ. (2004)
2. Faragó, I., Horváth, R.: Qualitative properties of monotone linear parabolic operators. In: E. J. Qualitative Theory of Diff. Equ., Proc. 8'th Coll. Qualitative Theory of Diff. Equ., vol. (8), pp. 1–15 (2008)
3. Horváth, R.: On the sign stability of numerical solutions of one-dimensional parabolic problem. Appl. Mat. Modell. 32(8), 1570–1578 (2008)
4. Krylov, N.V.: Nonlinear Elliptic and Parabolic Equations. D. Reidel, Dordrecht (1987)
5. Liu, H., Yong, J.: Option pricing with an illequid underlying asset market. J. of Econ. Dynamics and Control 29, 21–25 (2005)
6. Ortega, J., Rheinboldt, W.: Iterative Solution of Nonlinear Equations in Several Variables. Acad. Press, N.Y. (1970)
7. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker Inc., New York (2001)
8. Songzhe, L.: Existence of solution to initial value problem for a parabolic Monge-Ampère equation and application. Nonl. Anal. 65, 59–78 (2006)
9. Yong, J.: Introduction to mathematical finance. In: Yong, J., Cont, R. (eds.) Mathematical Finance-Theory and Applications. High Education Press, Beijing (2000)

Convergence of Finite Difference Schemes for a Multidimensional Boussinesq Equation

Natalia T. Kolkovska

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
acad. Bonchev str. bl.8, 1113 Sofia, Bulgaria
natali@math.bas.bg

Abstract. *Conservative finite difference schemes* for the numerical solution of multi-dimensional Boussinesq-type equations are constructed and studied theoretically. Depending on the way the nonlinear term $f(u)$ is approximated, two families of finite difference schemes are developed. Error estimates for these numerical methods in the uniform metric and the Sobolev space W_2^1 are obtained. The extensive numerical experiments given in [7] for the one-dimensional problem show good precision and full agreement between the theoretical results and practical evaluation for single soliton and the interaction between two solitons.

1 Introduction

1.1. Consider the Cauchy problem of the Boussinesq type equation (BE)

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= \Delta u + \beta_1 \Delta \frac{\partial^2 u}{\partial t^2} - \beta_2 \Delta^2 u + \alpha \Delta f(u), \quad x \in \mathbb{R}^d, \quad t > 0; \\ u(x, 0) &= u_0(x); \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x); \quad u(x, t) \rightarrow 0, \quad \Delta u(x, t) \rightarrow 0, \quad |x| \rightarrow \infty, \end{aligned} \quad (1)$$

where f is a smooth non-linear function, say $f(u) = u^2$, the amplitude parameter α is a real number and the dispersion parameters β_1 and β_2 are positive constants. BE (1) occurs in a number of mathematical models of real processes, for example, in the modeling of surface waves in shallow water. The essentials of the derivation of (1) from the full Boussinesq model can be found, e.g. in [3].

BE (1) called in [3] “Boussinesq Paradigm Equation” and similar BE, called “good BE”, “damped BE”, “improved BE”, “generalized double dispersion equation”, have been studied by many authors in the case of one dimensional (1D) space variable x (i.e. $d = 1$). The existence (both local and global in time) and uniqueness of weak and strong solutions in Sobolev spaces for the 1D problem are treated in [8,13,14]. Sufficient conditions for blow-up of the solution are given in [6,13]. Numerical solutions based on finite difference methods, spectral and pseudo-spectral methods and finite element methods can be found in [3,5,8,10,11].

The multidimensional version of BE (i.e. $d > 1$) is less studied. The dependence of existence, smoothness and blow-up of the solution on the nonlinear

function $f(u)$ is investigated in [14,15] for isotropic Sobolev spaces and in [12] for specially designed anisotropic Sobolev spaces. The numerical investigation of the 2D BE is also in its initial stage (see e.g. [1,2]).

In the present paper we study two families of finite difference schemes (FDS) for numerical computation of the multidimensional BE introduced in [7]. They differ on the way the approximation of the nonlinear term $\Delta f(u)$ is done. In Section 3 we show that one of the FDS retains an important property – the conservation law of the solution to the initial BE, while the other obeys a proper balance equation and demonstrates smaller approximation errors in experiments.

Section 4 contains error estimates for both FDS in the uniform metric and in the Sobolev space W_2^1 on the fixed time layer, as well as a number of corollaries and comments. The main results are contained in the convergence theorems 4 and 5. We establish second order of convergence for both FDS in the discrete W_2^1 norm, which is compatible with the rate of convergence of the similar linear problem. The convergence of both schemes (in the 1D case) is demonstrated in [7] on two basic examples of one solitary wave and interaction of two solitary waves traveling with different speeds towards each other. A variant of the proposed 2D FDS is implemented in [4].

Other FDS properties connected with the algorithms for their implementation can be found in [7]. Here we only mention that both FDS can be split as pairs of an elliptic and a hyperbolic 2D discrete equations, thus, their numerical solutions can be efficiently evaluated with stable algorithms.

1.2. By the linear change of variables $\frac{1}{\sqrt{\beta_1}}x = \xi$, $\frac{\sqrt{\beta_2}}{\beta_1}t = \theta$ equation (1) is rewritten in the form

$$\frac{\partial^2 U}{\partial \theta^2} = \Delta U + \Delta \frac{\partial^2 U}{\partial \theta^2} - \Delta^2 U + \Delta \frac{\beta_1}{\beta_2} \left(\alpha f(U) + \left(1 - \frac{\beta_2}{\beta_1} \right) U \right),$$

with $U(\xi, \theta) = u(x, t)$. Therefore, without loss of generality, we shall study the following problem

$$\frac{\partial^2 u}{\partial t^2} = \Delta u + \Delta \frac{\partial^2 u}{\partial t^2} - \Delta^2 u + \Delta g(u), \quad x \in \mathbb{R}^d, \quad t > 0, \quad (2)$$

$$u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x), \quad x \in \mathbb{R}^d, \quad (3)$$

$$u(x, t) \rightarrow 0, \quad \Delta u(x, t) \rightarrow 0, \quad |x| \rightarrow \infty, \quad t > 0, \quad (4)$$

where g is connected to f by

$$g(u) = \frac{\beta_1}{\beta_2} \left(\alpha f(u) + \left(1 - \frac{\beta_2}{\beta_1} \right) u \right).$$

We assume in this paper that the solution u to problem (2) – (4) belongs to $C^{6,4}(\mathbb{R}^d \times (0, T))$. Here $C^{m,n}(\mathbb{R}^d \times (0, T))$ denotes the space of continuous functions with continuous derivatives up to order m with respect to x and order n with respect to t . The existence of a classical (local or global) solution with the smoothness prescribed above is proved in the 1D case in [14], while for the multi-dimensional case similar results for *local* solutions are established in [15].

2 Numerical Method

The numerical methods described here work for any space dimension. For simplicity we present them in the case $d = 2$.

Let L_1, L_2 be sufficiently large numbers. We consider the discrete problem in the computational domain $\Omega = [-L_1, L_1] \times [-L_2, L_2]$, assuming that the solution with its derivatives is negligible outside this domain. We introduce a uniform grid with steps h_1, h_2 in Ω and let τ denote the uniform time step. The grid points are (x_i, y_j, t_k) , where $x_i = ih_1, i = -N_1, \dots, N_1; y_j = jh_2, j = -N_2, \dots, N_2; t_k = k\tau, k = 0, 1, 2, \dots$ with $N_1 = L_1/h_1, N_2 = L_2/h_2$. The discrete approximation to u at mesh point (x_i, y_j, t_k) is denoted by $v_{(i,j)}^{(k)}$. In the following, whenever possible, we omit the notation $v_{(i,j)}^{(k)}$ for the arguments of the mesh function v .

By the symbol C with different indexes we shall denote positive constants, which does not depend on parameters h, τ, γ, σ and on the functions u_0, u_1, g, u, v . By the symbol M with different indexes we shall denote positive constants, which depend on the norms of the functions u, v .

The standard 5-point discrete Laplacian is denoted by Δ_h . The finite difference approximation to the second time derivative is

$$v_{\bar{t}\bar{t},(i,j)}^{(k)} = \left(v_{(i,j)}^{(k+1)} - 2v_{(i,j)}^{(k)} + v_{(i,j)}^{(k-1)} \right) \tau^{-2}.$$

For a real parameter σ denote by v^σ the symmetric σ -weighted approximation to $v_{(i,j)}^{(k)}$ given by $v_{(i,j)}^{\sigma(k)} = \sigma v_{(i,j)}^{(k+1)} + (1-2\sigma)v_{(i,j)}^{(k)} + \sigma v_{(i,j)}^{(k-1)}$. We apply approximations with parameter σ to the purely spatial operators Δ_h and $(\Delta_h)^2$ in (2). The simplest way to approximate $g(v)$ at (x_i, y_j, t_k) is to take $g(v_{(i,j)}^{(k)})$. Thus, at interior grid points we obtain a first family of finite difference methods depending on the parameter σ

$$v_{\bar{t}\bar{t}} - \Delta_h v_{\bar{t}\bar{t}} - \Delta_h v^\sigma + (\Delta_h)^2 v^\sigma = \Delta_h g(v). \quad (5)$$

Another well known approximation to the nonlinear term at (x_i, y_j, t_k) is

$$g_1(v_{(i,j)}^{(k)}) = \frac{G(v_{(i,j)}^{(k+1)}) - G(v_{(i,j)}^{(k-1)})}{v_{(i,j)}^{(k+1)} - v_{(i,j)}^{(k-1)}}, \quad \text{where } G(u) = \int_0^u g(s) ds. \quad (6)$$

Note that in the classical case $f(u) = u^2$ the function g is a second degree polynomial and the anti-derivative G used in g_1 is explicitly evaluated. In this way we get the second family of finite difference schemes

$$v_{\bar{t}\bar{t}} - \Delta_h v_{\bar{t}\bar{t}} - \Delta_h v^\sigma + (\Delta_h)^2 v^\sigma = \Delta_h g_1(v). \quad (7)$$

An $O(|h|^2 + \tau^2)$ approximation to the initial conditions (3) is given by

$$v_{(i,j)}^{(0)} = u_0(x_i, y_j), \quad (8)$$

$$v_{(i,j)}^{(1)} = u_0(x_i, y_j) + \tau u_1(x_i, y_j) + 0.5 \tau^2 (I - \Delta_h)^{-1} (\Delta_h u_0 - (\Delta_h)^2 u_0 + \Delta_h g(u_0))(x_i, y_j). \quad (9)$$

For the approximation of the second boundary condition the mesh is extended outside the domain Ω by one line at each space boundary and the symmetric second-order finite difference is used for the approximation of the second spatial derivative in (4).

Equations (5) or (7) with initial conditions (8), (9) and boundary conditions described above form two families of finite difference schemes indexed by σ . The efficient algorithms for evaluation of their solutions are given in [7].

3 Discrete Identities

For given time moment t_k we consider the space of mesh functions $v^{(k)}$ which vanish at the points on the boundary of Ω and we define the operator $A = -\Delta_h$. In this space denote by $\langle v^{(k)}, w^{(k)} \rangle = \sum_{i,j} h_1 h_2 v_{(i,j)}^{(k)} w_{(i,j)}^{(k)}$ the discrete scalar product of mesh functions $v^{(k)}, w^{(k)}$ with respect to the spatial variables.

In the space of functions, which satisfy both asymptotic conditions on the computational boundary (2) we define the operator $B = (I + A)(I + \sigma\tau^2 A)$. Note that A and B are self-adjoint positive definite operators.

For the analysis of difference schemes, we use the representation $v^\sigma = v + \sigma\tau^2 v_{\bar{t}t}$ and rewrite the equations (5) and (7) in the operator form

$$Bv_{\bar{t}t} + Av + A^2v = -Ag, \quad (10)$$

$$Bv_{\bar{t}t} + Av + A^2v = -Ag_1. \quad (11)$$

Following [7], we first define the functional E_h^L given by

$$\begin{aligned} (E_h^L v)^{(k)} &= \left\langle A^{-1/2} v_t^{(k)}, A^{-1/2} v_t^{(k)} \right\rangle + \tau^2 (\sigma - 1/4) \left\langle (I + A) v_t^{(k)}, v_t^{(k)} \right\rangle \\ &\quad + \left\langle v_t^{(k)}, v_t^{(k)} \right\rangle + 1/4 \left\langle v^{(k)} + v^{(k+1)} + A(v^{(k)} + v^{(k+1)}), v^{(k)} + v^{(k+1)} \right\rangle. \end{aligned}$$

and then, by incorporating the non-linear term g_1 , the full discrete “energy” functional

$$(E_h v)^{(k)} = (E_h^L v)^{(k)} + \left\langle G(v^{(k+1)}), 1 \right\rangle + \left\langle G(v^{(k)}), 1 \right\rangle.$$

The following theorems are proved in [7]:

Theorem 1 (Discrete conservation law). *The discrete “energy” $(E_h v)^{(k)}$ of the solution v to the scheme (11) is preserved in time, i.e. it satisfies the equalities*

$$(E_h v)^{(k)} = (E_h v)^{(0)}, \quad k = 1, 2, \dots \quad (12)$$

The discrete balance law (12) valid for the solution to the scheme (11) fully corresponds to the energy equation [14] valid for the solution to the initial problem (2)–(4). The scheme (10) does not have a strict conservation of the discretized energy functional $(E_h v)^{(k)}$, but it satisfies similar *balance identities* given below.

Theorem 2. *The solution to the scheme (10) satisfies the equalities*

$$(E_h^L v)^{(k)} - (E_h^L v)^{(k-1)} + \left\langle g(v^k), v^{(k+1)} - v^{(k-1)} \right\rangle = 0, \quad k = 1, 2, \dots \quad (13)$$

4 Convergence of the FDS

4.1 Analysis of the Linear Problem

We begin with the analysis of the following discrete linear problem

$$Bv_{\bar{t}t} + Av + A^2v = -A\psi_1 + \psi_2, \quad (14)$$

where ψ_1 and ψ_2 are given functions. The initial conditions to (14) are (8) and (9) with v_0, v_1 on the place of u_0, u_1 and $-A\psi_1 + \psi_2$ on the place of $-Ag(u_0)$.

Using the stability theory from [9], Chapter 6, we get the following theorem:

Theorem 3. *Let γ be a positive real number. Assume that for some steps h and τ the parameter σ satisfies the inequality*

$$\sigma > \frac{1+\gamma}{4} - \frac{1}{\tau^2 \|A\|}. \quad (15)$$

Then the finite difference method (14), (8), (9) is stable with respect to the initial data and the right-hand side. Moreover, the following estimate holds:

$$\begin{aligned} (v^{(k)}, v^{(k)}) + (Av^{(k)}, v^{(k)}) &\leq C \frac{1+\gamma}{\gamma} \left[(Bv^{(0)}, v^{(0)}) + (A^{-1}Bv_t^{(0)}, A^{-1}Bv_t^{(0)}) \right. \\ &\quad \left. + \sum_{s=1}^{k-1} \tau (\psi_1^{(s)}, \psi_1^{(s)}) + \sum_{s=1}^{k-1} \tau (A^{-1}\psi_2^{(s)}, A^{-1}\psi_2^{(s)}) \right]. \quad (16) \end{aligned}$$

4.2 Convergence of the FDS's for the Non-linear Problem

Now we are ready to study the convergence of FDS. We begin with FDS (10) assuming for the smoothness of the non-linear term $g \in W_\infty^1(\mathbb{R})$. Denote by $z = v - u$ the error of the solution. We substitute $v = z + u$ into the problem (10) and obtain the following problem for the error z :

$$Bz_{\bar{t}t} + Az + A^2z = -Ag(v) - Bu_{\bar{t}t} - Au - A^2u. \quad (17)$$

Now we use the equation (2) and Taylor series for the function u about the node (x_i, y_j, t_k) . It is straightforward to show that

$$-Ag(v) - Bu_{\bar{t}t} - Au - A^2u = -A\psi_1 + \psi_2$$

with $\psi_1 = g(v) - g(u)$, $\psi_2 = O(|h|^2 + \tau^2)$. Thus, we get that (17) has the form of (14) and we can apply Theorem 3. We estimate ψ_1 by $|g(v^{(k)}) - g(u(t_k))| \leq M^{(k)}|z^{(k)}|$ with a constant $M^{(k)}$ chosen so that $\max_{i,j} (|u(x_i, y_j, t_k)|, |v_{i,j}^{(k)}|) \leq M^{(k)}$.

Also (8) and (9) approximate the initial conditions (3) locally with $O(|h|^2 + \tau^2)$ error. In this way we get

$$\begin{aligned} (z^{(k)}, z^{(k)}) + (Az^{(k)}, z^{(k)}) &\leq C \frac{1+\gamma}{\gamma} \left(C_1(|h|^2 + \tau^2)^2 + \sum_{s=1}^{k-1} \tau M^{(s)} (z^{(s)}, z^{(s)}) \right). \quad (18) \end{aligned}$$

Proceeding by induction on k if we assume the boundedness of $z^{(s)}$ for $s = 1, 2, \dots, k-1$ we shall obtain from (18) that $|z_{i,j}^{(k)}|$ is bounded and hence $|v_{i,j}^{(k)}|$ is bounded whenever $\|u(\cdot, \cdot, t_k)\|$ is bounded.

Now we use the Gronwall's lemma and conclude

$$\left(z^{(k)}, z^{(k)}\right) + \left(Az^{(k)}, z^{(k)}\right) \leq \frac{1+\gamma}{\gamma} C e^{Mt_k} (|h|^2 + \tau^2)^2 \quad (19)$$

with $M = \max_k M^{(k)}$. In this way we proved the following theorem

Theorem 4. Assume $g \in W_\infty^1(\mathbb{R})$, the parameter σ satisfies (15) for some $\gamma > 0$ and the solution u to the problem (2) – (4) obey $u \in C^{6,4}(\mathbb{R}^2 \times (0, T))$. Then the solution v to the finite difference scheme (10), (8), (9) converges to u as $|h|, \tau \rightarrow 0$ and the estimate (19) holds for the error $z = y - u$ of the scheme.

Now we turn to FDS (11) assuming for the smoothness of the non-linear term $g \in W_\infty^2(\mathbb{R})$. We may use the same arguments as in the previous scheme, but taking into account that ψ_1 is different. Here

$$\psi_1^{(s)} = \frac{G(v^{(s+1)}) - G(v^{(s-1)})}{v^{(s+1)} - v^{(s-1)}} - g(u(t_s)).$$

We first expand $G(v^{(s+1)})$ in Taylor series about the point $v^{(s-1)}$ and then we expand $g(v^{(s-1)}) = G'(v^{(s-1)})$ in Taylor series about the point $u(t_s)$. Thus, we get

$$|\psi_1^{(s)}| < C \left(M_1^{(s)} \tau^2 + M_2^{(s)} \left(|z^{(s-1)}| + |z^{(s)}| + |z^{(s+1)}| \right) \right),$$

where $M_2^{(s)}$ is a constant satisfying

$$M_2^{(s)} \geq \max_{i,j} \left(\left| u(x_i, y_j, t_s) \right|, \left| \frac{\partial^2 u}{\partial t^2}(x_i, y_j, t_s) \right|, |v_{i,j}^{(s-1)}|, |v_{i,j}^{(s)}|, |v_{i,j}^{(s+1)}| \right).$$

Now Theorem 3 gives

$$\left(z^{(k)}, z^{(k)}\right) + \left(Az^{(k)}, z^{(k)}\right) \leq C_2 \frac{1+\gamma}{\gamma} \left(C_1 (|h|^2 + \tau^2)^2 + \sum_{s=1}^k \tau M_2^{(s)} \left(z^{(s)}, z^{(s)} \right) \right).$$

The above inequality differs from (18) by the term containing $(z^{(k)}, z^{(k)})$ in the right-hand side. If τ is sufficiently small, say $\tau \leq 0.5\gamma \left(C_2(1+\gamma)M_2^{(s)} \right)^{-1}$, then this term can be moved to the left-hand side and, thus, we see that $z^{(k)}$ satisfies (18) (with a bigger constant C). Using once more the Gronwall's lemma we obtain the following result:

Theorem 5. Assume $g \in W_\infty^2(\mathbb{R})$ and the parameter σ satisfies (15) with some $\gamma > 0$. Assume that the solution u to (2) – (4) obeys $u \in C^{6,4}(\mathbb{R}^2 \times (0, T))$ and

the solution v to the finite difference scheme (11), (8), (9) is bounded in the maximal norm. Let M be a constant such that

$$M \geq \max_{i,j,s} \left(|u(x_i, y_j, t_s)|, \left| \frac{\partial^2 u}{\partial t^2}(x_i, y_j, t_s) \right|, |v_{i,j}^{(s)}| \right)$$

and τ be sufficiently small, $\tau < \gamma (C_2(1 + \gamma)M)^{-1}$. Then v converges to the exact solution u as $|h|, \tau \rightarrow 0$ and the following estimate holds for the error $z = y - u$:

$$\left(z^{(k)}, z^{(k)} \right) + \left(Az^{(k)}, z^{(k)} \right) \leq \frac{1 + \gamma}{\gamma} C e^{Mt_k} (|h|^2 + \tau^2)^2. \quad (20)$$

The assumption in Theorem 5 for boundedness of the discrete solution could be dropped. It can be derived from the other assumptions by proving separately that the iterative process for obtaining $v^{(k+1)}$ from (11) is convergent. The proof uses that some mappings are contractive as in [14]. Here we skip the proof due to its length. We underline that the other difference between Theorems 4 and 5 – the hypothesis for the upper estimate on τ in Theorem 5 – is essential.

4.3 Corollaries

The main feature of Theorems 4 and 5 is the established second order of convergence in discrete W_2^1 norm, which is compatible with the rate of convergence of the similar linear problem.

Corollary 1. (i) The convergence of the solution to FDS (10) or FDS (11) with $\sigma > 0.25$ to the exact solution is of second order when $|h|$ and τ go independently to zero.

(ii) The convergence of the solution to the explicit FDS (10) or FDS (11) with $\sigma = 0$ to the exact solution is of second order when $|h|$ and τ go to 0 provided: $\tau < \frac{|h|}{\sqrt{1+\gamma}}$ for the 1D problem or $\tau < \frac{|h|}{\sqrt{2(1+\gamma)}}$ for the 2D case.

The error estimates obtained in Theorems 4 and 5 are in the discrete W_2^1 norm on the $t^{(k)}$ time layer. Using embedding theorems for the uniform norm we derive

Corollary 2. Under the assumptions of Theorems 4 or 5 the FDS (10) or (11) admits the following error estimate in the uniform norm:

$$\begin{aligned} \max_i |z_i^{(k)}| &< C e^{Mt_k} \sqrt{\frac{1 + \gamma}{\gamma}} (|h|^2 + \tau^2), \quad d = 1; \\ \max_{i,j} |z_{i,j}^{(k)}| &< C e^{Mt_k} \sqrt{\ln N} \sqrt{\frac{1 + \gamma}{\gamma}} (|h|^2 + \tau^2), \quad d = 2. \end{aligned}$$

The above estimates are optimal for the 1D case and *almost* optimal (up to a logarithmic factor) for the 2D case.

One of the main assumptions in Theorems 4 and 5 is the boundedness of the exact solution u to the BE on the time interval $[0, T]$. Such assumption is

natural because the BE may have both bounded on the time interval $[0, \infty)$ solutions and blowing up solutions. The L_∞ norm of the solution is included in the exponent in the right-hand sides of the error estimates in Theorems 4 and 5. Hence, if u blows up at a moment T_0 which is slightly bigger than T , then $\|u\|_{L_\infty[0, T]}$ will be big and, hence, the term e^{MT} will be big and the convergence will slow up. Additional, but not so important restriction on the time step τ , is the upper bound in Theorem 5 containing the reciprocal of $\|u\|_{L_\infty[0, T]}$. In any case the FDS should be applied with very small τ 's if one would like to evaluate the solution in a neighborhood of the blow up moment.

Acknowledgments. The author is grateful to C. I. Christov for the numerous valuable discussions.

References

1. Chertock, A., Christov, C.I., Kurganov, A.: Central-Upwind Schemes for the Boussinesq paradigm Equation (submitted)
2. Christou, M., Christov, C.I.: Galerkin Spectral Method for the 2D Solitary Waves of Boussinesq Paradigm Equation. In: AIP, vol. 1186, pp. 217–224 (2009)
3. Christov, C.I.: An energy-consistent dispersive shallow-water model. *Wave Motion* 34, 161–174 (2001)
4. Christov, C.I., Kolkovska, N., Vasileva, D.: On the Numerical Simulation of Unsteady Solutions for the 2D Boussinesq Paradigm Equation. LNCS, vol. 6046. Springer, Heidelberg (to appear)
5. Christov, C.I., Velarde, M.: Inelastic Interaction of Boussinesq Solutions. *Intern. J. Bifurcation Chaos* 4, 1095–1112 (1994)
6. Liu, Y., Xu, R.: Potential well method for Cauchy problem of generalized double dispersion equations. *J. Math. Anal. Appl.* 338, 1169–1187 (2008)
7. Kolkovska, N.: Two Families of Finite Difference Schemes for the Multidimensional Boussinesq Equation. In: AIP (to appear)
8. Pani, A., Saranga, H.: Finite Element Galerkin Method for the “Good” Boussinesq Equation. *Nonlinear Analysis* 29, 937–956 (1997)
9. Samarsky, A.: *The Theory of Difference Schemes*. Marcel Dekker Inc., New York (2001)
10. Ortega, T., Sanz-Serna, J.M.: Nonlinear stability and convergence of finite-difference methods for the “good” Boussinesq equation. *Numer. Math.* 58, 215–229 (1990)
11. El-Zoheiry: Numerical study of the improved Boussinesq equation. *Chaos, Solitons and Fractals* 14, 377–384 (2002)
12. Varlamov, V.: Two-dimensional Boussinesq equation in a disc and anisotropic Sobolev spaces. *C. R. Mecanique* 335, 548–558 (2007)
13. Wang, S., Chen, G.: The Cauchy Problem for the Generalized IMBq Equation in $W^{s,p}(R^n)$. *J. Math. Anal. and Appl.* 266, 38–54 (2002)
14. Wang, S., Chen, G.: Cauchy problem of the generalized double dispersion equation. *Nonlinear Analysis* 64, 159–173 (2006)
15. Xu, R., Liu, Y.: Global existence and nonexistence of solution for Cauchy problem of multidimensional double dispersion equations. *J. Math. Anal. Applic.* 359, 729–751 (2009)

A Numerical Approach for Obtaining Fragility Curves in Seismic Structural Mechanics: A Bridge Case of Egnatia Motorway in Northern Greece

Asterios Liolios¹, Panagiotis Panetsos², Angelos Liolios¹,
George Hatzigeorgiou³, and Stefan Radev⁴

¹ Democritus University of Thrace, Department of Civil Engineering,
Institute of Structural Mechanics and Earthquake Engineering, Xanthi, Greece
`liolios@civil.duth.gr`

² Egnatia Odos S.A., Bridge Maintenance Department, Themi-Thessaloniki, Greece

³ Democritus University of Thrace, Department of Environmental Engineering,
Lab. Ecological Mechanics and Technology, Xanthi, Greece

⁴ Bulgarian Academy of Sciences, Institute of Mechanics,
Acad. G. Bonchev Str., Bl. 4, 1113 Sofia, Bulgaria
`stradev@imbm.bas.bg`

Abstract. Fragility curves for Civil Engineering structures represent a critically important step in seismic damage estimation process. In the present article, a numerical methodology for the evaluation of such curves for bridges is presented. The methodology is based on the Finite Element Method, combines the nonlinear static pushover procedure with the capacity spectrum method and is applied for establishing fragility curves for an existing reinforced concrete bridge with seismic stoppers in the Krystalopigi - Psilorahi section of Egnatia Motorway, in the county of Epirus, northern Greece.

Keywords: Computational Earthquake Engineering, Fragility Curves of Bridges.

1 Introduction

As well known [1], the key element in formulating mitigation and disaster planning strategies in Earthquake Engineering is the realistic estimation of the urban seismic risk. In this respect, development of vulnerability relationships for both, the existing and under design Civil Engineering structures, represents a critically important step in damage estimation process. Scope of the vulnerability analysis is the creation of the so-called fragility curves [1]–[4],[9]–[11], through which the probability that a specific damage level will be exceeded for a given intensity of a seismic event may be quickly estimated, supporting significantly the decision-making procedures. So, fragility curves for Civil Engineering Structures, such as buildings and especially bridges, are a useful tool for the assessment of the

damage they may sustain for a certain level of earthquake shaking. In combination with seismic hazard analysis at the bridge sites, they can lead to a reliable assessment of the seismic risk of highways. Furthermore, they can even be used by the authorities in charge to prioritize the on site aftershock inspections, in order to check the structural integrity of the bridges subjected to a severe seismic event.

Several methodologies dealing with the assessment of fragility curves for bridges can be found in recent literature, based on either empirical or analytical procedures [2], [3],[9]–[11]. Also, methodologies originally proposed for buildings can sometimes be extended for use in the case of bridges [2]–[4],[14].

In the present article, a numerical methodology for the evaluation of vulnerability curves for bridges having deck on precast beams, seating through elastomeric bearings on the piers and with seismic stoppers is presented. The methodology is based on the Finite Element Method and combines the nonlinear static pushover procedure and the capacity spectrum method [1]–[4],[9]–[11]. The methodology is applied for establishing fragility curves for an existing reinforced concrete bridge crossing a steep slope in the Kristallopigi - Psilorahi section of Egnatia Motorway, in the county of Epirus, Northern Greece.

Egnatia Odos is a new motorway that crosses Northern Greece in an E-W direction. It is currently the largest and technically the most demanding highway project in Greece, and one of the biggest ones under current (2008-2009) construction in Europe. Moreover, for the design and construction of Egnatia Motorway, a lot of Applied Mechanics topics are involved, e.g. structural and seismic mechanics, geotechnical and transport engineering, hydraulic and environmental engineering, etc. So, Egnatia Motorway can be considered as an active field of Applied Mechanics. Its main axis has a length of 670 km and includes about 1900 special structures (bridges, tunnels and culverts). These structures are expected to withstand several minor or moderate earthquakes during their life, and may be damaged if they are subjected to a major (catastrophic) earthquake. So, the construction of their fragility curves is very significant. The bridge examined herein is a structurally representative one of many bridges in Egnatia Motorway, and in Greece more generally.

2 Methods for Assessing Structural Vulnerability

The vulnerability functions, required for the fragility curves, are expressed [2]–[4], [9]–[11] in terms of a Lognormal cumulative probability function in the form of next eq. (1):

$$P_f(DP \geq DP_i|S) = \Phi \left[\frac{1}{\beta_{tot}} \cdot \ln \left(\frac{S}{S_{mi}} \right) \right] \quad (1)$$

Here $P_f(\cdot)$ is the probability of the damage parameter DP being at, or exceeding, the value DP_i for the i -th damage state for a given seismic intensity level defined by the earthquake parameter S (here the Peak Ground Acceleration-PGA or Spectral Displacement- S_d), Φ is the standard cumulative probability function,

S_{mi} is the median threshold value of the earthquake parameter S required to cause the i -th damage state, and β_{tot} is the total lognormal standard deviation. Thus, the description of the fragility curve involves the two parameters, S_{mi} and β_{tot} , which must be determined.

Now we consider briefly the problem of computing the vulnerability functions (1) for Civil Engineering Structures, such as buildings and especially bridges. For the latter ones, the case of reinforced concrete bridges with seismic stoppers is herein investigated. This case is a contact mechanics problem. So, such bridges can be considered as nonlinear elastic and inelastic systems with impacts which arise in mechanical and civil engineering applications. In Civil Engineering applications, such systems arise also, besides in the above analysis of bridges with seismic stoppers, in the analysis of pounding of adjacent buildings.

Next it is briefly described the general problem of the seismic pounding of adjacent structures. This problem belongs to the so-called Dynamic Inequality Problems of Mechanics, for which a strict mathematical treatment can be obtained by using the variational or hemivariational inequality concept. As well known, the latter one has been introduced in Mechanics by P.D. Panagiotopoulos [5]. As concerns their numerical treatment, many significant contributions are already available, see e.g. [5], [6]. So, for the case of two interacting structures (A) and (B), following e.g. the procedure of [7], the problem is first formulated as an inequality one by using concepts of Non-Convex Analysis. Next, double discretization, in space by the Finite Element Method and in time by a direct-time integration scheme (e.g. the central difference method), and optimization methods are used. Thus, by piecewise linearization of the interface unilateral contact laws, at each time-step a nonconvex linear complementarity problem of the following matrix form with reduced number of unknowns is finally solved:

$$\mathbf{v} \geq \mathbf{0}, \quad \mathbf{A}\mathbf{v} + \mathbf{a} \leq \mathbf{0}, \quad \mathbf{v}^T \cdot (\mathbf{A}\mathbf{v} + \mathbf{a}) = \mathbf{0}. \quad (2)$$

So, the nonlinear Response Time-History (RTH) for a given seismic ground excitation can be computed.

As was mentioned in the Introduction, the present study focuses on the simplified practical fragility analysis of bridges, that involve impacts due to the seismic stoppers designed to effectively withstand earthquake loads and reduce the size of the piers. For such a practical simplified analysis, these systems are represented by single and multi degree of freedom models with piecewise linear elastic stiffness elements that often involve strong inelastic behavior in parts of the system. So, the previous general approach for pounding of adjacent structures is simplified by considering the simple bridge with seismic stoppers shown in Figure 1a. The bridge deck is connected to the piers by elastomeric bearings and seismic stoppers are added on the pier caps that have a small gap with the deck structure so that the elastomeric bearings are free to move under ambient or traffic loads, while they impact on the stoppers only under moderate or strong earthquake loads. Activation of the stoppers due to impact results in

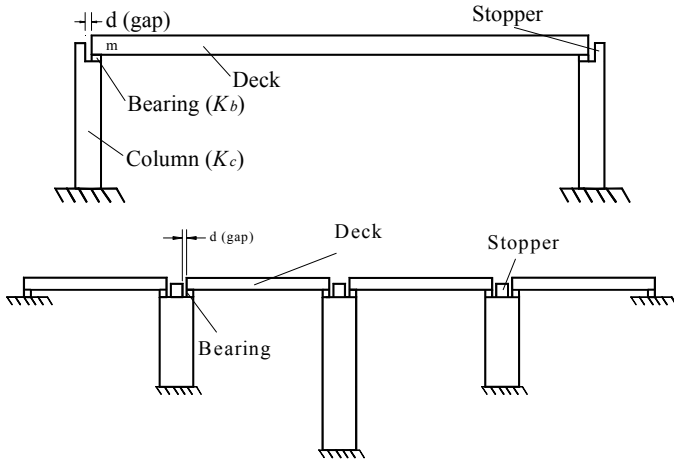


Fig. 1. Schematic diagram of: (a) single span bridge (top); (b) multi span bridge (bottom)

sudden increase of the stiffness of the structure. The gaps between the stoppers and the bearings are usually selected such that the impact with the stoppers occurs before the pier yielding.

From the previous analysis is obvious that the damage level depends on the input seismic excitation, i.e. the seismic ground acceleration. As well known from Structural Dynamics and Earthquake Engineering [1], because this input is not known for future earthquakes, the *spectral approach* is used according to various aseismic building codes, e.g. the Greek Aseismic Code EAK2000 [12]. So here, instead of a non-linear dynamic analysis, which is time consuming [1], the approach of [4], [14] is followed. According to equation (1), the description of the fragility curve involves only two parameters, S_{mi} and β_{tot} . The first parameter S_{mi} is estimated on the basis of the capacity spectrum method [1], wherein the demand spectrum is plotted for a range of values of the earthquake parameter S (in spectral acceleration vs. spectral displacement format) and it is superimposed on the same plot with the capacity curve of the bridge. The earthquake parameter used in this study is the peak ground acceleration (PGA). The second parameter of Eq. (1) is the total lognormal standard deviation β_{tot} , which takes into account the uncertainties in seismic input motion (demand), in the response and resistance of the bridge (capacity), and in the definition of damage states. This parameter (β_{tot}) can be estimated by a statistical combination of the individual uncertainties (in demand, capacity, and damage state definition) assuming these are statistically independent. On the basis of empirical fragility curves obtained from actual bridge damage data, the value of β_{tot} was set in [4],[14] equal to 0.6; due to the lack of a more accurate estimation of uncertainties in capacity, demand and damage states. Briefly, the proposed methodology comprises the following main steps:

- (a) Due to elastomeric bearings, the system of the deck and prestressed reinforced concrete (r/c) beams is moving horizontally up to the existed gaps of spans will close. Here, the shear stiffness of the system of elastomeric bearings is quite active.
- (b) A Finite Element Model of the bridge is constructed using linear elements and lumped plastic hinges, for the end sections of the piers, the bents, the continuity slabs and the abutment's ballast walls.
- (c) The structural elements possess suitable effective flexural stiffness.
- (d) The structural critical sections are analyzed in order to calculate the bilinear moment-curvature (M-C) diagram, as well as the moment-axial force diagram up to the yielding point by using a suitable material law for confined concrete.
- (e) Transformations of bilinear diagrams M-C in bilinear diagrams M-R (moments-rotations) using a suitable length of each plastic hinge.
- (f) The first translational mode-shape distribution of external static seismic lateral forces is considered in the nonlinear static pushover analysis, for both horizontal principal axes, which represent adequately the dynamic response of the bridge.
- (g) The gravity loads of the system are in action.
- (h) Static pushover procedure and capacity spectrum method are performed.
- (i) The damage levels of the bridge are defined and finally the statistical log-normal function of probability distribution is used.

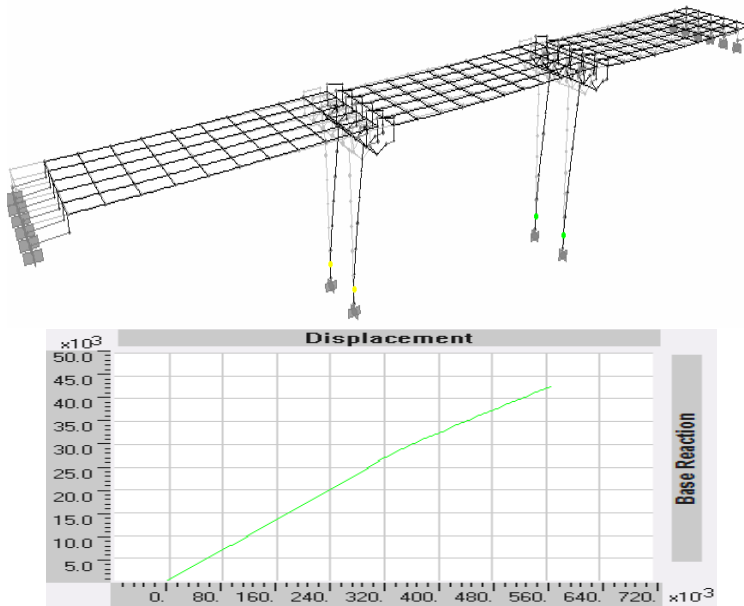


Fig. 2. G2 bridge, longitudinal direction : FEM model (top); pushover curve (no gap closure) (bottom)

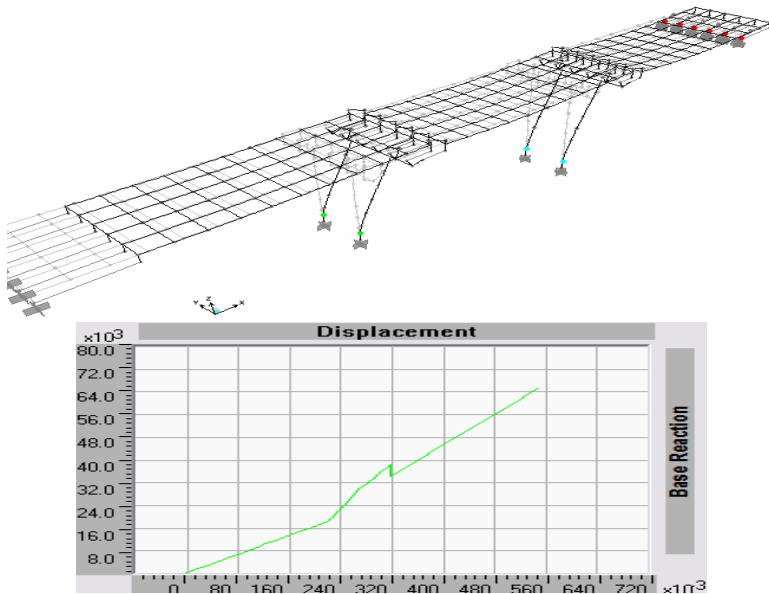


Fig. 3. G2 bridge, longitudinal direction : FEM model (top); Pushover curve considering gap closure of the end expansion joints as well as inelastic response of the abutment - backfill complex (bottom)

3 The Case of an Egnatia Motorway Bridge with Seismic Stoppers

The bridge considered herein is the G2 valley-bridge near Kristallopigi, Epirus, built on the west sector of the Egnatia Motorway, in northern Greece. The 100m long bridge is carrying the right branch of the motorway over a steep mountainous slope near Kristallopigi. The bridge consists of three equal spans, each constructed using six 33m long prestressed - precast concrete beams that rest on two piers and two abutments via elastomeric bearings.

The reinforced concrete piers are twin square columns, 20m high, framed by an orthogonal beam that supports the precast beams through 6 type NB4 rectangular elastomeric bearings with dimensions 600x700x255 (135) in (mm). A 25cm thick in situ reinforced concrete slab, on the top of the beams, continues over the piers. It is acting as a diaphragm along the total length of the bridge, which is separated by the abutment ballast walls through elastometallic anchored joints, by gaps of 20 cm. Stoppers on the pier's beams were designed to be distant from the superstructure such as to be activated after the exceeding of the maximum spectral displacement.

Details for the geometric and elastic characteristics of the bridge elements are given in [8], [14], where also the computation steps for obtaining the fragility

Table 1. Definition of damage states

i	Damage state	Necessary repair interventions	Duration of interventions	Damage ratio $D_i = \delta_i/\delta_y$
0	No damage	None	- - -	< 0.7
1	Minor damage	Small-scale repairs	< 3 days	> 0.7
2	Moderate damage	Repair of structural elements	< 3 weeks	> 1.5
3	Extensive damage	Reconstruction of structural parts	< 3 monts	> 3
4	Collapse	Reconstruction of bridge	> 3 monts	μ_u

curves are given in details. Herein we refer briefly to Figs 2 and 3, which show the Finite Element Modelling by using the SAP2000 program [13] for the modal pushover analyses, and to the Table 1 concerning five states of damage. Table 1 concerns five states of damage ($i=0$ to 4), which were defined as a function of the damage ratio $D = \delta/\delta_\gamma$, where δ is the displacement at the target point and δ_γ the corresponding yield displacement. Corresponding threshold values D_i that define the boundaries between the damage states were also defined. Finally, Figure 4 and Figure 5 shows the fragility curves, which were computed assuming a lognormal cumulative probability distribution for the damage ratio as a function of peak ground acceleration PGA. A first interpretation of these analytically derived curves leads to the conclusion that the longitudinal direction is more critical, as having bigger probability of failure.

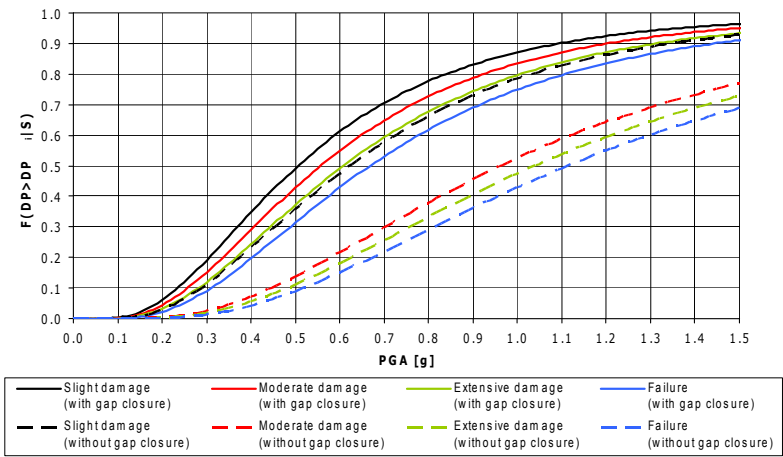


Fig. 4. Fragility curves of the G2 Kristallopigi bridge: Longitudinal direction

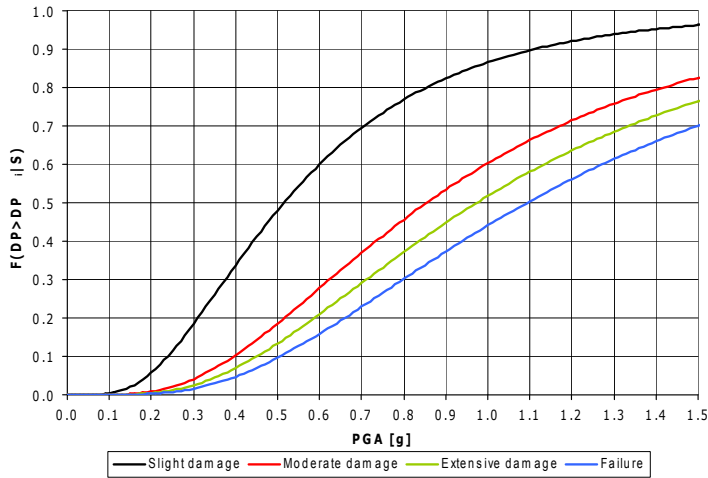


Fig. 5. Fragility curves of the G2 Kristallopigi bridge: Transverse direction

4 Conclusions

A simplified numerical methodology has been presented for the calculation of the vulnerability curves of bridges in the presence of seismic stoppers. This methodology is based on the Finite Element Method, on a modal pushover nonlinear static analysis and on a capacity demand spectrum approach, instead of a time consuming non-linear dynamic based vulnerability analysis. Using the aforementioned approach, fragility curves were developed for the G2 Kristallopigi valley bridge of Egnatia Motorway, Northern Greece.

References

1. Chopra, A.K.: Dynamics of Structures. Theory and Applications to Earthquake Engineering. Pearson Prentice Hall, New Jersey (2007)
2. Elnashai, A., Rossetto, T.: Derivation of Vulnerability Functions for European Type RC Structures Based on Observational Data. *Engineering Structures* 25, 1241–1263 (2003)
3. Shinozuka, M., Feng, M.Q., Lee, J., Naganuma, T.: Statistical Analysis of Fragility Curves. *Journal of Engineering Mechanics* 126(12), 1224–1231 (2000)
4. Makarios, T., Lekidis, V., Kappos, A., Karakostas, C., Moschonas, J.: Development of seismic vulnerability curves for a bridge with elastomeric bearings. In: Papadrakakis, M., et al. (eds.) *Proceedings of the COMPDYN 2007, ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering*, Rethymno, Crete, Greece, June 13–16 (2007)
5. Panagiotopoulos, P.D.: *Hemivariational Inequalities and Applications in Mechanics and Engineering*. Springer, Berlin (1993)
6. Panagiotopoulos, P.D., Glocker, C.: Inequality constraints with elastic impacts in deformable bodies. The convex case. *Arch. Appl. Mech.* 70, 349–365 (2000)

7. Liolios, A.A.: A linear complementarity approach to the nonconvex dynamic problem of unilateral contact with friction between adjacent structures. *Z. Angew. Math. Mech. (ZAMM)* 69, T420–T422 (1989)
8. Liolios, A., Panetsos, P., Makarios, T.: Seismic fragility functions for a bridge of Egnatia motorway in northern Greece. In: *Proceedings of 6th German-Greek-Polish Symposium "Recent Advances in Mechanics"*, Alexandroupolis, Greece, September 17–21 (2007)
9. Hwang, H.H.M., Jaw, J.W.: Probabilistic damage analysis of structures. *J. struct. Enging. ASCE* 116(7), 1992–2007 (1990)
10. Shinozuka, M., Hwang, H., Reich, M.: Reliability assessment of reinforced concrete containment structures. *Nuc. Enging. Des.* 80, 247–267 (1984)
11. Park, Y.-J., Ang, A.H.-S.: Mechanistic Seismic Damage Model for Reinforced Concrete. *Journal of Structural Engineering (ASCE)* 111, 740–757 (1985)
12. EAK 2000: Greek Aseismic Code. Ministry of Public Works and Environment, OASP (Organization of Seismic Protection), Athens (2000)
13. SAP 2000: Linear and Non linear Static and Dynamic Analysis and Design of Three-Dimensional Structures. Computers and Structures Inc., Berkeley, California (2005)
14. ASPROGE: Research Project for the ASeismic Protection of Bridges. Egnatia Odos S.A., Thessaloniki, Greece (2007)

An Efficient Numerical Method for a System of Singularly Perturbed Semilinear Reaction-Diffusion Equations

S. Chandra Sekhara Rao and Sunil Kumar

Department of Mathematics, Indian Institute of Technology Delhi,
Hauz Khas, New Delhi-110 016, India
scsr@maths.iitd.ac.in, skumar.iitd@gmail.com

Abstract. In this work we consider a system of singularly perturbed semilinear reaction-diffusion equations. To solve this problem numerically, we construct a finite difference scheme of Hermite type, and combine this with standard central difference scheme in a special way on a piecewise-uniform Shishkin mesh. We prove that the method is third order uniformly convergent. Numerical experiments are conducted to demonstrate the efficiency of the present method.

Keywords: Singularly perturbed, System of semilinear equations, Shishkin mesh, Parameter-uniform convergence.

1 Introduction

In this work we develop an efficient numerical method for solving a system of singularly perturbed semilinear reaction-diffusion equations. These systems of equations arise for example in catalytic reaction theory [1]. The simplified physical problem involves an isothermal reaction which is catalyzed in a pellet. Scalar singularly perturbed semilinear problems have been extensively studied in the literature, see [2–5] and the references therein. However, the study of systems of singularly perturbed semilinear equations is limited. These problems were solved asymptotically in [6, 7] and numerically in [8, 9]. It is well known that classical numerical methods are not appropriate for singularly perturbed problems. Therefore, various non-classical approaches are used to design special numerical methods that converge uniformly no matter how small the perturbation parameter ε , see [2, 10]. One of the most attractive approaches is to use standard finite difference schemes on specially designed meshes.

We consider the following system of semilinear equations

$$\overline{\mathbf{T}}\mathbf{u} := -\mathbf{E}\mathbf{u}'' + \mathbf{f}(x, \mathbf{u}) = \mathbf{0}, \quad x \in \Omega = (0, 1), \quad \mathbf{u}(0) = \mathbf{p}, \quad \mathbf{u}(1) = \mathbf{q}, \quad (1)$$

where $\mathbf{E} = \text{diag}(\varepsilon, \dots, \varepsilon)$ with small parameter $0 < \varepsilon \leq 1$, $\mathbf{u} = (u_1, \dots, u_M)^T$, and $\mathbf{f}(x, \mathbf{u}) = (f_1(x, \mathbf{u}), \dots, f_M(x, \mathbf{u}))^T$ is a sufficiently smooth vector function.

We assume that, for all $(x, \mathbf{y}) \in \overline{\Omega} \times \mathbb{R}^M$

$$\frac{\partial f_k}{\partial u_i}(x, \mathbf{y}) \leq 0, \quad k \neq i, \quad \text{and} \quad \sum_{i=1}^M \frac{\partial f_k}{\partial u_i}(x, \mathbf{y}) > \alpha^* > 0, \quad k = 1, \dots, M. \quad (2)$$

Assumption (2) and the implicit function theorem ensures the existence of a unique solution $\mathbf{u}$ of (1) and also that of the associated reduced problem $\mathbf{f}(x, \mathbf{u}_0) = \mathbf{0}$, for all $x \in \Omega$, defined by setting $\varepsilon = 0$ in the differential equation in (1).

Methods of high order convergence reduce computational cost to find good numerical approximations. In this paper, we construct a finite difference scheme of Hermite type, and combine this with standard central difference scheme in a special way on a piecewise-uniform Shishkin mesh to solve the system numerically. Error analysis is given and parameter-uniform error bounds are established.

This paper is arranged as follows. In section 2, we discretize the problem (1) on a piecewise-uniform Shishkin mesh. In section 3, error analysis is given and parameter-uniform error bounds are established. Results of numerical experiments are presented in section 4.

Notations: Throughout the paper, we use C to denote a generic positive constant independent of ε and the discretization parameter. Similarly, $\mathbf{C} = (C, C, \dots, C)^T$ is a vector of identical constants with the same independencies. Define $\mathbf{v} \leq \mathbf{w}$ if $v_k \leq w_k$, $1 \leq k \leq M$, and $|\mathbf{v}| = (|v_1|, \dots, |v_M|)^T$. For any function $g \in C(\overline{\Omega})$, define $g_j = g(x_j)$; if $\mathbf{g} \in C(\overline{\Omega})^M$ then $\mathbf{g}_j = \mathbf{g}(x_j) = (g_{1;j}, \dots, g_{M;j})^T$. For a closed and bounded set D , $\|\mathbf{g}\|_D$ is the maximum norm of $\mathbf{g}$ and $\|\mathbf{g}\|_D = \max\{\|\mathbf{g}_1\|_D, \dots, \|\mathbf{g}_M\|_D\}$. If $D = \overline{\Omega}$, we consider usual notation $\|\cdot\|_\infty$.

2 Discretization

We first define a piecewise-uniform Shishkin mesh $\overline{\Omega}^N := \{x_i\}_0^N$. Let $N = 2^k$, $k \geq 2$ be a positive integer. Define the transition parameter

$$\sigma = \min \left\{ \frac{1}{4}, \quad \sigma_0 \sqrt{\varepsilon} \ln N \right\}, \quad (3)$$

where σ_0 is a positive constant to be fixed later. We divide $\overline{\Omega}$ into three subintervals $[0, \sigma]$, $[\sigma, 1 - \sigma]$, and $[1 - \sigma, 1]$, where the subintervals $[0, \sigma]$ and $[1 - \sigma, 1]$ represent the inner regions and the subinterval $[\sigma, 1 - \sigma]$ represents the outer region. The subinterval $[0, \sigma]$ and $[1 - \sigma, 1]$ is divided into $\frac{N}{4}$ equidistant elements and the subinterval $[\sigma, 1 - \sigma]$ is divided into $\frac{N}{2}$ equidistant elements.

Set $i_0 = \frac{N}{4}$, then $x_{i_0} = \sigma$ and $x_{N-i_0} = 1 - \sigma$ are the transition points. Let $x_i = x_{i-1} + h_i, \forall i = 1, \dots, N$. Then the resulting piecewise-uniform mesh is represented as

$$h_i = \begin{cases} h = \frac{4\sigma}{N}, & \text{for } i = 1, \dots, i_0; \\ H = \frac{2(1-2\sigma)}{N}, & \text{for } i = i_0 + 1, \dots, N - i_0; \\ h = \frac{4\sigma}{N}, & \text{for } i = N - i_0 + 1, \dots, N. \end{cases} \quad (4)$$

Note that if $\sigma = 1/4$, then the mesh is uniform, N^{-1} is very small with respect to ε and therefore a classical analysis could be used to prove the uniform convergence of the scheme. So, here we only consider the case $\sigma = \sigma_0 \sqrt{\varepsilon} \ln N$.

Following the construction made in [4] for the scalar singularly perturbed semilinear problem, we construct a finite difference scheme of Hermite type for the system of singularly perturbed semilinear equations (1). We call this the fourth order Hermite scheme. We consider a special combination of the fourth order Hermite scheme and the central difference scheme on a piecewise-uniform Shishkin mesh $\overline{\Omega}^N$ to discretize the system of singularly perturbed semilinear equations (1). The discrete operator $\mathbf{T} = (T_1, \dots, T_M)^T$ is defined as

$$\mathbf{T} \mathbf{U}_j = \mathbf{0} \quad \text{for } j = 1, \dots, N-1, \quad \mathbf{U}(0) = \mathbf{p}, \quad \mathbf{U}(1) = \mathbf{q}, \quad (5)$$

where

$$\begin{aligned} T_k \mathbf{U}_j := & r_j^{k,-} U_{k,j-1} + r_j^{k,c} U_{k,j} + r_j^{k,+} U_{k,j+1} + q_j^{k,-} f_k(x_{j-1}, \mathbf{U}_{j-1}) \\ & + q_j^{k,c} f_k(x_j, \mathbf{U}_j) + q_j^{k,+} f_k(x_{j+1}, \mathbf{U}_{j+1}), \quad k = 1, \dots, M, \quad j = 1, \dots, N-1. \end{aligned}$$

The coefficients $r_j^{k,\star}$, $\star = -, c, +$, are given by

$$r_j^{k,-} = \frac{-2\varepsilon}{h_j(h_j + h_{j+1})}, \quad r_j^{k,c} = \frac{2\varepsilon}{h_j h_{j+1}}, \quad r_j^{k,+} = \frac{-2\varepsilon}{h_{j+1}(h_j + h_{j+1})}. \quad (6)$$

The coefficients $q_j^{k,\star}$, $\star = -, c, +$, are defined in two different ways.

(i) For $j = 1, \dots, i_0 - 1, N - i_0 + 1, \dots, N - 1$, i.e., $x_j \in (0, \sigma) \cup (1 - \sigma, 1)$, the coefficients $q_j^{k,\star}$, $k = 1, \dots, M$, $\star = -, c, +$, are given by

$$\begin{aligned} q_j^{k,-} &= \frac{h_j^2 - h_{j+1}^2 + h_j h_{j+1}}{6h_j(h_j + h_{j+1})}, & q_j^{k,c} &= \frac{h_j^2 + h_{j+1}^2 + 3h_j h_{j+1}}{6h_j h_{j+1}}, \\ q_j^{k,+} &= \frac{h_{j+1}^2 - h_j^2 + h_j h_{j+1}}{6h_{j+1}(h_j + h_{j+1})}. \end{aligned} \quad (7)$$

(ii) For $j = i_0, \dots, N - i_0$, i.e., $x_j \in [\sigma, 1 - \sigma]$, the coefficients $q_j^{k,\star}$, $k = 1, \dots, M$, $\star = -, c, +$, are defined in two different cases. Let $\frac{\partial f_k}{\partial u_k}(x, \mathbf{y}) \leq \beta_{kk}$, for all $(x, \mathbf{y}) \in \overline{\Omega} \times \mathbb{R}^M$.

First, if $2H^2 \beta_{kk}/3 \leq \varepsilon$, the coefficients $q_j^{k,\star}$, $j = i_0 + 1, \dots, N - i_0 - 1$, $k = 1, \dots, M$, $\star = -, c, +$, are defined again by (7). For $j = i_0, N - i_0$, i.e., for the transition points, the coefficients $q_j^{k,\star}$, $k = 1, \dots, M$, $\star = -, c, +$, are given by

$$q_j^{k,-} = 1/3, \quad q_j^{k,c} = 1/3, \quad q_j^{k,+} = 1/3. \quad (8)$$

In the other case, if $2H^2 \beta_{kk}/3 > \varepsilon$, the coefficients $q_j^{k,\star}$, $j = i_0, \dots, N - i_0$, $k = 1, \dots, M$, $\star = -, c, +$, are given by

$$q_j^{k,-} = 0, \quad q_j^{k,c} = 1, \quad q_j^{k,+} = 0. \quad (9)$$

Note that, we considered the fourth order Hermite scheme in the boundary layer region $(0, \sigma) \cup (1 - \sigma, 1)$. While in the regular region $[\sigma, 1 - \sigma]$, we considered the central difference scheme, if $2H^2 \beta_{kk}/3 > \varepsilon$. In the other case, if $2H^2 \beta_{kk}/3 \leq \varepsilon$, we considered the fourth order Hermite scheme in $(\sigma, 1 - \sigma)$ and a slightly modified scheme at the transition points σ and $1 - \sigma$. This modification at the transition points is considered, since, in general, at the transition points, the coefficients defined by (7) are not positive and thus the Frechet-derivative $\mathbf{T}'$ of $\mathbf{T}$ is not an M-matrix.

Lemma 1. *Let N_0 be the smallest positive integer such that*

$$4\sigma_0^2 \max_{1 \leq k \leq M} \{\beta_{kk}\}/3 < N_0^2 / \ln^2 N_0.$$

Then, for any $N \geq N_0$, the Frechet-derivative $\mathbf{T}'$ satisfies

$$\|\mathbf{T}'^{-1}\|_\infty \leq \frac{1}{\min\{1, \alpha^*\}}. \quad (10)$$

Proof. From (2) and (6)-(9), it immediately follows that the Frechet-derivative $\mathbf{T}'$ is an M-matrix with all of its rows satisfying

$$(T')_{ii} - \sum_{i \neq j} |(T')_{ij}| \geq \min\{1, \alpha^*\} > 0.$$

Then from the Theorem A of Varga [11], (10) follows.

An immediate consequence of above lemma is that the discrete operator $\mathbf{T}$ satisfies the comparison principle and it is parameter-uniform stable in the maximum norm.

3 Convergence Analysis

In this section, we investigate the accuracy of the present method. For the analysis we need sharp bounds on the exact solution $\mathbf{u}$ of (1) and its derivatives. An application of the technique in [5] gives the following result.

Lemma 2. *Let $\mathbf{u}$ be the solution of the problem (1). Let $\alpha \in (0, \alpha^*)$ be arbitrary but fixed. Then*

$$|\mathbf{u}^{(m)}(x)| \leq C(1 + \varepsilon^{-m/2}(e^{-x\sqrt{\alpha/\varepsilon}} + e^{-(1-x)\sqrt{\alpha/\varepsilon}})), \quad (11)$$

for all $x \in \overline{\Omega}$ and $m = 0, \dots, 6$.

Theorem 1. *Let $\mathbf{u}$ be the solution of the problem (1) and $\mathbf{U}$ that of the problem (5) on a piecewise-uniform Shishkin mesh. Then, for any $N \geq N_0$,*

$$\|\mathbf{u} - \mathbf{U}\|_\infty \leq C(N^{-3} + N^{-4} \ln^4 N). \quad (12)$$

Proof. We write the k th component of the truncation error

$$T_k \mathbf{u}(x_j) = \phi_j^k u_k(x_j) \quad \text{for } j = 1, \dots, N-1,$$

where

$$\begin{aligned} \phi_j^k u_k(x_j) &= r_j^{k,-} u_k(x_{j-1}) + r_j^{k,c} u_k(x_j) + r_j^{k,+} u_k(x_{j+1}) \\ &\quad + \varepsilon q_j^{k,-} u_k''(x_{j-1}) + \varepsilon q_j^{k,c} u_k''(x_j) + \varepsilon q_j^{k,+} u_k''(x_{j+1}). \end{aligned}$$

We estimate the truncation error of the present method in the following cases.

(i). For $x_j \in (0, \sigma) \cup (1 - \sigma, 1)$, we have $h_j = h_{j+1} = 4\sigma_0 \sqrt{\varepsilon} N^{-1} \ln N$. Then Taylor expansions give

$$|T_k \mathbf{u}(x_j)| \leq C \varepsilon h_j^4 \|u_k^{(6)}\|_{[x_{j-1}, x_{j+1}]}, \quad k = 1, \dots, M.$$

Now use $h_j = 4\sigma_0 \sqrt{\varepsilon} N^{-1} \ln N$ and $\|\mathbf{u}^{(6)}\|_\infty \leq C \varepsilon^{-3}$ to get

$$|T_k \mathbf{u}(x_j)| \leq C(\sigma_0 N^{-1} \ln N)^4 \quad \text{for } x_j \in (0, \sigma) \cup (1 - \sigma, 1), k = 1, \dots, M.$$

(ii). For $x_j \in [\sigma, 1 - \sigma]$, we need a special decomposition of the exact solution $\mathbf{u}$ into regular part $\mathbf{v}$ and layer part $\mathbf{w}$. Set $x^* = 4\sqrt{\varepsilon} \alpha^{-1/2} \ln(1/\sqrt{\varepsilon})$ and define for each $k \in \{1, \dots, M\}$ and $x \in \overline{\mathcal{Q}}$

$$v_k(x) = \begin{cases} \sum_{\ell=0}^6 \frac{(x - x^*)^\ell}{\ell!} u_k^{(\ell)}(x^*) & \text{for } 0 \leq x \leq x^*; \\ u_k(x) & \text{for } x^* \leq x \leq 1 - x^*; \\ \sum_{\ell=0}^6 \frac{(x - x^*)^\ell}{\ell!} u_k^{(\ell)}(1 - x^*) & \text{for } 1 - x^* \leq x \leq 1, \end{cases}$$

and $w_k(x) = u_k(x) - v_k(x)$. Then the Lemma 2 and the choice of x^* yields

$$|v_k^{(m)}(x)| \leq C(1 + \varepsilon^{2-m/2}) \quad \text{and} \quad (13)$$

$$|w_k^{(m)}(x)| \leq C \varepsilon^{-m/2} (e^{-x\sqrt{\alpha/\varepsilon}} + e^{-(1-x)\sqrt{\alpha/\varepsilon}}) \quad \text{for } m = 0, \dots, 6, \quad (14)$$

cf. [12]. Here we consider two distinct cases.

(iia). For the case $2H^2 \beta_{kk}/3 > \varepsilon$, central difference scheme is used. Then for $\mathbf{g} \in C^4(\overline{\mathcal{Q}})^M$, by Taylor expansions

$$|\phi_j^k g_k(x_j)| \leq \begin{cases} C \varepsilon \|g_k^{(2)}\|_{[x_{j-1}, x_{j+1}]}, \\ C \varepsilon (h_j + h_{j+1}) \|g_k^{(3)}\|_{[x_{j-1}, x_{j+1}]}, \\ C \varepsilon h_j^2 \|g_k^{(4)}\|_{[x_{j-1}, x_{j+1}]}, \quad \text{if } h_j = h_{j+1}. \end{cases} \quad (15)$$

Using decomposition of $\mathbf{u}$, we write

$$|\phi_j^k u_k(x_j)| \leq |\phi_j^k v_k(x_j)| + |\phi_j^k w_k(x_j)|.$$

For bounding the truncation error in $\mathbf{v}$, we use last two estimates of (15). For the layer part $\mathbf{w}$, we use first estimate of (15). This yields

$$|\phi_j^k u_k(x_j)| \leq \begin{cases} C\varepsilon N^{-1} & \text{if } x_j \in \{\sigma, 1 - \sigma\} \\ C\varepsilon N^{-2} & \text{if } x_j \in (\sigma, 1 - \sigma) \end{cases} + \max_{x \in [x_{j-1}, x_{j+1}]} |e^{-x\sqrt{\alpha/\varepsilon}} + e^{-(1-x)\sqrt{\alpha/\varepsilon}}|.$$

Choose $\sigma_0 \geq 4\alpha^{-1/2}$ and use $2H^2 \beta_{kk}/3 > \varepsilon$; this leads to

$$|\phi_j^k u_k(x_j)| \leq N^{-3} \quad \text{for } j = i_0, \dots, N - i_0.$$

Collecting various bounds, we get

$$|T_k \mathbf{u}(x_j)| \leq C(N^{-4} \ln^4 N + N^{-3}) \quad \text{for } 2H^2 \beta_{kk}/3 > \varepsilon, x_j \in \Omega^N. \quad (16)$$

(iib). Now consider the case $2H^2 \beta_{kk}/3 \leq \varepsilon$. Analogous to the decomposition of $\mathbf{u}$, we decompose truncation error

$$|\phi_j^k u_k(x_j)| \leq |\phi_j^k v_k(x_j)| + |\phi_j^k w_k(x_j)|.$$

For the regular part $\mathbf{v}$, Taylor expansions give

$$|\phi_j^k v_k(x_j)| \leq \begin{cases} C\varepsilon(h_j + h_{j+1})^2 \|v_k^{(4)}\|_{[x_{j-1}, x_{j+1}]} & \text{if } x_j \in \{\sigma, 1 - \sigma\}; \\ C\varepsilon h_j^4 \|v_k^{(6)}\|_{[x_{j-1}, x_{j+1}]} & \text{if } x_j \in (\sigma, 1 - \sigma). \end{cases}$$

Using (13), for $k = 1, \dots, M$, we get

$$|\phi_j^k v_k(x_j)| \leq \begin{cases} C\varepsilon N^{-2} & \text{if } x_j \in \{\sigma, 1 - \sigma\}; \\ CN^{-4} & \text{if } x_j \in (\sigma, 1 - \sigma). \end{cases}$$

For the layer part $\mathbf{w}$, by Taylor expansions and (14), we get

$$|\phi_j^k w_k(x_j)| \leq C\varepsilon \|w_k^{(2)}\|_{[x_{j-1}, x_{j+1}]} \leq CN^{-4} \quad \text{for } j = i_0, \dots, N - i_0.$$

Collecting various bounds, for $2H^2 \beta_{kk}/3 \leq \varepsilon$, $x_j \in \Omega^N$, we get

$$|T_k \mathbf{u}(x_j)| \leq CN^{-4} \ln^4 N + \begin{cases} C\varepsilon N^{-2} & \text{if } x_j \in \{\sigma, 1 - \sigma\}, \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

To improve these bounds, we use the barrier function technique. Defining the barrier function $\mathbf{Z}^\pm(x_j) = \pm C(N^{-3} \sqrt{\varepsilon} \ln N \theta(x_j) + N^{-4} \ln^4 N + N^{-3}) + \mathbf{U}(x_j)$, where θ is the piecewise linear polynomial

$$\theta(x) := \begin{cases} x/\sigma & \text{for } x \in [0, \sigma], \\ 1 & \text{for } x \in [\sigma, 1 - \sigma], \\ (1 - x)/\sigma & \text{for } x \in [1 - \sigma, 1], \end{cases}$$

using the comparison principle for the operator $\mathbf{T}$, it follows that

$$|(\mathbf{u} - \mathbf{U})(x_j)| \leq C(N^{-3} \sqrt{\varepsilon} \ln N + N^{-4} \ln^4 N + N^{-3}),$$

and taking into account that $\sqrt{\varepsilon} \ln N \leq C$ (for $\sigma < 1/4$) the result follows.

4 Numerical Results

To demonstrate the efficiency of the present method, we consider the following test problem [9]

$$-\varepsilon u_1'' + u_1 - 1 - (1 - u_1)^3 + \exp(u_1 - u_2) = 0, \quad u_1(0) = u_1(1) = 0,$$

$$-\varepsilon u_2'' + u_2 - 0.5 - (0.5 - u_2)^5 + \exp(u_2 - u_1) = 0, \quad u_2(0) = u_2(1) = 0.$$

To solve the corresponding nonlinear system of equations associated with the discrete problem, the Newton's method is used with zero as the initial guess. The stopping criterion is $\|\mathbf{U}^{(k)} - \mathbf{U}^{(k-1)}\|_\infty < 10^{-15}$. Here $\mathbf{U}^{(k)}$, for $k = 1, 2, \dots$, represent the successive approximates to $\mathbf{U}$ computed iteratively.

We consider $\alpha = 0.99$ in the construction of piecewise-uniform Shishkin mesh $\overline{\Omega}^N$. The exact solution of the test example is not known. We use the double mesh method to compute the numerical rate of convergence. To do this, we compute not only $\mathbf{U}$, but also another approximate solution $\widetilde{\mathbf{U}}$ to the problem (1) on the mesh $\widetilde{\overline{\Omega}}^N$ with a slightly altered mesh parameter $\widetilde{\sigma}$, where

$$\widetilde{\sigma} = \min \left\{ \frac{1}{4}, \quad \sigma_0 \sqrt{\varepsilon} \ln(N/2) \right\}.$$

Here the altered mesh parameter is used such that the i th mesh point of the mesh $\overline{\Omega}^N$ coincides with the $(2i)$ th mesh point of the mesh $\widetilde{\overline{\Omega}}^{2N}$. We compute the maximum errors E_ε^N and the parameter-uniform errors E^N by

$$E_\varepsilon^N = \max_{0 \leq j \leq N} |(\mathbf{U})_j - (\widetilde{\mathbf{U}})_{2j}| \quad \text{and} \quad E^N = \max_\varepsilon E_\varepsilon^N.$$

Table 1. Maximum errors and numerical rates of convergence of the present method for the Example 1

$\varepsilon = 2^{-k}$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
$k=4$	8.06E-08	5.05E-09	3.16E-10	1.97E-11	1.24E-12
	4.00	4.00	4.00	4.00	
8	1.90E-05	1.28E-06	8.03E-08	5.03E-09	3.15E-10
	3.89	3.99	4.00	4.00	
12	1.20E-03	2.00E-04	1.90E-05	1.28E-06	8.03E-08
	2.59	3.40	3.89	3.99	
16	1.20E-03	2.00E-04	2.20E-05	2.34E-06	2.28E-07
	2.59	3.18	3.23	3.36	
20	1.20E-03	2.00E-04	2.20E-05	2.34E-06	2.28E-07
	2.59	3.18	3.23	3.36	
24	1.20E-03	2.00E-04	2.20E-05	2.34E-06	2.28E-07
	2.59	3.18	3.23	3.36	
28	1.20E-03	2.00E-04	2.20E-05	2.34E-06	2.28E-07
	2.59	3.18	3.23	3.36	
E^N	1.20E-03	2.00E-04	2.20E-05	2.34E-06	2.28E-07
R^N	2.59	3.18	3.23	3.36	

The numerical rates r_ε^N and the parameter-uniform numerical rates r^N are calculated by

$$R_\varepsilon^N = \ln(E_\varepsilon^N / E_\varepsilon^{2N}) / \ln(2) \quad \text{and} \quad R^N = \ln(E^N / E^{2N}) / \ln(2).$$

For the different values of ε and N , the maximum errors E_ε^N and the numerical rates R_ε^N of the present method applied to the test problem is given in Table 1. The last two rows in the table represents the parameter-uniform errors E^N and the parameter-uniform convergence rates R^N . Numerical results given in Table 1 clearly supports the theoretical estimates established in previous section.

References

1. Chang, K.W., Howes, F.A.: Nonlinear Singular Perturbation Phenomena. Springer, New York (1984)
2. Roos, H.-G., Stynes, M., Tobiska, L.: Robust numerical methods for singularly perturbed differential equations. Springer, Berlin (2008)
3. Surla, K., Uzelac, Z.: A uniformly accurate spline collocation method for a normalized flux. J. Comput. Appl. Math. 166, 291–305 (2004)
4. Herceg, D.: Uniform fourth order difference scheme for a singular perturbation problem. Numer. Math. 56, 675–694 (1990)
5. Vulcanovic, R.: On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh. Zb. Rad. Prirod. Mat. Fak. Univ. Novom Sadu Ser. Mat. 13, 187–201 (1983)
6. Jeffries, J.S.: A singularly perturbed semilinear system. Meth. Appl. Anal. 3, 157–173 (1996)
7. Zong-chi, L., Su-rong, L.: Singularly perturbed phenomena of semilinear second order systems. Appl. Math. Mech. 9, 1131–1138 (1988)
8. Shishkina, L., Shishkin, G.I.: Conservative Numerical Method for a System of Semilinear Singularly Perturbed Parabolic Reaction-Diffusion Equations. Math. Modell. Anal. 14, 211–228 (2009)
9. Gracia, J.L., Lisbona, F.J., Madaune-Tort, M., O’Riordan, E.: A system of singularly perturbed semilinear equations. In: Hegarty, A., Kopteva, N., O’Riordan, E., Stynes, M. (eds.). Lect. Notes Comput. Sci. Eng., vol. 69, pp. 163–172 (2009)
10. Miller, J.J.H., O’Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems. World Scientific, Singapore (1996)
11. Varga, R.S.: On diagonal dominance arguments for bounding $\|A^{-1}\|_\infty$. Linear Algebra Appl. 14, 211–217 (1976)
12. Linss, T.: The necessity of Shishkin-decompositions. Appl. Math. Lett. 14, 891–896 (2001)

A Comparison of Methods for Solving Parametric Interval Linear Systems with General Dependencies

Iwona Skalna

AGH University of Science and Technology, Krakow, Poland
skalna@agh.edu.pl

Abstract. This study compares two methods for solving interval linear systems whose coefficients are functions of interval parameters: the generalized Rump's fixed-point iteration and Skalna's Direct Method. Both methods have the same scope of application and require estimating the range of the same functions over a box. Evaluation of functional ranges using the simplest form of interval analysis produces wide intervals. This is due in a large part to the so-called interval dependency. To cope with the dependence problem, revised affine arithmetic with a new affine approximation of a product is used. Numerical examples are provided to show the advantages of Skalna's Direct Method over generalized Rump's fixed point iteration.

1 Introduction

The problem of solving parametric linear systems is of great importance in many real-life problems, which are very often subject to uncertainty. The latter can be caused by many factors (e.g. approximation of model structure or model parameters, numerical approximations) and there are many ways of dealing with it. When uncertainty is modelled using interval numbers, then instead of a parametric linear system, a family of parametric linear systems known as the *parametric interval linear system* (PILS), is considered.

Several methods for solving PILS have been developed in recent years, see e.g. [1], [3], [5], [11], [17], [19], [20], [21]. It seems, however, that the fixed point iteration developed by Rump [16,17], studied in e.g. [15], improved independently by Popova [14] and Skalna, and used in e.g. [2], [11], [13], is the best known method for solving PILS.

In this study, it is argued that a less widely known method for solving PILS, namely Skalna's Direct Method [21], has advantages over Rump's, because of being faster and less sensitive to the amount of uncertainty. Skalna's method is similar to Rump's fixed point iteration in that it has the same scope of application and it requires estimating the range of the same functions over a box. Evaluation of functional ranges using the simplest form of interval analysis often leads to overestimation. This is due in a large part to the so-called *interval dependency*. To cope with the dependency problem, revised affine arithmetic with a new affine approximation of a product is used.

The paper has the following structure. Basic facts on revised affine arithmetic and a new formula for multiplication of revised affine forms are presented in Section 2. Section 3 is devoted to the problem of solving parametric interval linear systems. Selected methods for solving such systems are discussed in Section 4. Section 5 contains numerical examples used to compare the methods from Section 4. The paper ends with concluding remarks.

2 Revised Affine Arithmetic

Revised affine arithmetic (RAA) (see e.g. [4], [6], [7], [9]) keeps track of correlations between quantities, therefore it is able to provide much tighter intervals than conventional interval arithmetic, especially in long computations. In revised affine arithmetic, a partially unknown quantity x is represented by an affine form $\hat{x} = x_0 + x_1\varepsilon_1 + \dots + x_n\varepsilon_n + e_x[-1, 1]$, which consists of two parts: a first degree polynomial of length n and the cumulative error $e_x[-1, 1]$ ($e_x > 0$ is an error variable), which represents the errors introduced by performing non-affine operations. The central value x_0 , the coefficients x_i (called partial deviations), and e_x are finite floating-point numbers, and $\varepsilon_i \in [-1, 1]$ are dummy variables. In rigorous computations, last term is also used to accumulate the rounding errors in floating-point arithmetic.

All the standard arithmetic operations as well as other classical functions are redefined for revised affine forms. Affine arithmetic operations result straightforwardly in affine forms. Extending non-affine operations requires using good affine approximation to the exact result. Below, a new affine approximation for a product of affine forms is suggested.

2.1 Multiplication of Affine Forms

The product of affine forms is a quadratic polynomial. It must be approximated by an affine form. Based on [10], the new affine approximation of the product is suggested: $\hat{z} = \hat{x}\hat{y} = z_0 + \sum_{i=1}^n z_i\varepsilon_i + e_z[-1, 1]$, where $z_0 = 2x_0y_0 + (\underline{d} + \overline{d})/2$, $z_i = x_0y_i + y_0x_i$, $e_z = (\overline{d} - \underline{d})/2 + e_x(|x_0| + \dots + |x_n|) + e_y(|y_0| + \dots + |y_n|) + e_x e_y$, $\underline{d}$ and $\overline{d}$ are, respectively, minimum and maximum of the quadratic term over $\langle \hat{x}, \hat{y} \rangle$. The algorithm for computing $\underline{d}$ and $\overline{d}$ is presented below.

```

for  $i = 1$  to  $n$  do
   $\alpha_x = \alpha_y = 0$ ;
  for  $j = 1$  to  $n$  do
    if  $x_i = 0$  then  $(x_j \geq 0)? e = 1 : e = -1$ ;
    else if  $y_i = 0$  then  $(y_j \geq 0)? e = 1 : e = -1$ ;
    else if  $-y_i/x_i \cdot x_j + y_j \geq 0$  then  $e = 1$ ; else  $e = -1$ ;
     $\alpha_x = \alpha_x + x_j \cdot e$ ;  $\alpha_y = \alpha_y + y_j \cdot e$ ;
  end
   $a = x_i \cdot y_i$ ;  $b = \alpha_x \cdot y_i + \alpha_y \cdot x_i$ ;  $c = \alpha_x \cdot \alpha_y - x_0 \cdot y_0$ ;
   $\underline{d} = \min \{ \underline{d}, \min \{ ax^2 + bx + c \mid x \in [-1, 1] \} \}$ 
   $\overline{d} = \max \{ \overline{d}, \max \{ ax^2 + bx + c \mid x \in [-1, 1] \} \}$ 
end

```

Algorithm 1. Computing $\underline{d}$ and $\overline{d}$

3 Parametric Interval Linear Systems

Consider a linear algebraic system $A(p)x(p) = b(p)$, where $p \in R^k$ is a vector of parameters, $A(p)$ is an $n \times n$ matrix, $b(p)$ is an n -dimensional vector, and $A_{ij}(p)$ and $b_i(p)$ ($i, j = 1, \dots, n$) are assumed to be continuous functions of parameters. When the parameters are considered to be unknown (or uncertain) and vary within prescribed intervals $p_i \in \mathbf{p}_i$, $i = 1, \dots, k$, a family of parametric linear system is obtained:

$$A(p)x(p) = b(p), \quad p \in \mathbf{p} \quad , \quad (1)$$

and is called *parametric interval linear system*. The corresponding non-parametric interval matrix and vector are denoted, respectively, by $A(\mathbf{p}) := \square\{A(p) \mid p \in \mathbf{p}\}$ and $b(\mathbf{p}) := \square\{b(p) \mid p \in \mathbf{p}\}$. Here, $\square$ denotes a *hull* which is defined for any bounded set S as $\square S = \bigcap \{\mathbf{Y} \in \mathbb{I}^n \mid S \subseteq \mathbf{Y}\} = [\inf S, \sup S]$.

The set of all solutions to (1), called *parametric (united) solution set*, is defined as:

$$S(A(p), b(p), \mathbf{p}) := \{x(p) \mid A(p)x(p) = b(p), \text{ for some } p \in \mathbf{p}\} \quad . \quad (2)$$

The solution set is bounded if $A(\mathbf{p})$ is regular; that is $A(p)$ is non-singular for every $p \in \mathbf{p}$. The hull $\square S(A(p), b(p), \mathbf{p})$ of the bounded parametric solution set is called an *interval hull solution*. It is quite expensive to obtain the solution set itself or its interval hull. In the general case, the problem of computing the hull solution is NP-hard. Therefore, an interval vector $\mathbf{x}^* \supseteq S(\mathbf{p}) \supseteq S(A(p), b(p), \mathbf{p})$, called the *outer interval solution*, is computed instead, and the goal is for $\mathbf{x}^*$ to be as narrow as possible.

4 Methods for Solving Parametric Interval Linear Systems

Two competing methods for solving PILS are presented below: Skalna's Direct Method and generalized Rump's fixed point iteration. Both methods have the same scope of application: Rump's method requires strong regularity of $A(\mathbf{p})$ [14], and Skalna's method requires $\square\{(\text{mid } A(\mathbf{p}))^{-1}A(p) \mid p \in \mathbf{p}\}$ to be an H-matrix [21]. It can be shown easily that those requirements are equivalent. Moreover, both methods require sharp bounds for the ranges of the following functions:

$$Z(p) = R \cdot (b(p) - A(p)\tilde{x}) \quad , \quad (3)$$

$$D(p) = R \cdot A(p) \quad . \quad (4)$$

on the domain $\mathbf{p} \in \mathbb{I}^k$, in order to obtain sharp parametric solution enclosure. In this study, revised affine arithmetic is used for bounding the ranges of (3) and (4). In the implementation, $R \approx (\text{mid } A(\mathbf{p}))^{-1}$ and $\tilde{x} \approx R \cdot \text{mid } b(\mathbf{p})$.

4.1 Rump's Fixed-Point Iteration

S. Rump [16] proposed the inclusion theorem which led to the fixed-point iteration method for the solution to an interval linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$. In [17], he gave a straightforward generalization to affine-linear dependencies in the matrix and the right hand side. A modification of Rump's method, which led to generalized Rump's fixed point iteration, was proposed independently by Popova [14] and Skalna. This modification consisted of computing $C(\mathbf{p})$ instead of $\mathbf{C} = \mathbf{I} - \mathbf{R}\mathbf{A}(\mathbf{p})$ (for details see [14]). Rump's method requires that $A(\mathbf{p})$ is strongly regular [14] or equivalently that $\rho(C(\mathbf{p})) < 1$ [15]. The pseudo-code of the generalized Rump's method (GRM) is presented below.

```

 $\tilde{x} = R \cdot \text{mid}(b(\mathbf{p}));$ 
 $C(\mathbf{p}) = \square\{I - RA(p) \mid p \in \mathbf{p}\} = I - \square\{RA(p) \mid p \in \mathbf{p}\} = I - D(\mathbf{p});$ 
 $Z(\mathbf{p}) = \square\{R \cdot (b(p) - A(p)\tilde{x}) \mid p \in \mathbf{p}\};$ 
 $\mathbf{V} = Z(\mathbf{p});$ 
repeat
   $\mathbf{Y} = \mathbf{V} \cdot [1 - \varepsilon, 1 + \varepsilon] + [-\mu, \mu];$ 
  for  $i = 1$  to  $n$  do
     $\mathbf{V}_i = Z(\mathbf{p}) + C(\mathbf{p}) \cdot (\mathbf{V}_1, \dots, \mathbf{V}_{i-1}, \mathbf{Y}_i, \dots, \mathbf{Y}_n)^T;$ 
  end
until  $\mathbf{V} \subset \mathbf{Y};$ 
return  $\tilde{x} + \mathbf{V};$ 

```

Algorithm 2. Generalized Rump's fixed point iteration

The *inflation parameter* ε is assumed to vary within the prescribed interval $(0, 1)$, and each component of μ should be equal to the smallest positive floating-point number. The results of the GRM depends on the problem to solve and on the choice of ε ([15], [18]). To maintain good relative accuracy, a small ε have to be chosen [17]. This, however, results in a larger number of iterations and, thus, longer computation time. On the other hand, by increasing ε the number of the iterations in the GRM can be decreased; however, at the expense of accuracy.

4.2 Skalna's Direct Method

This single-step Direct Method for solving parametric interval linear system with general dependencies was proposed in [21]. The pseudo-code of Skalna's Direct Method (SDM) is given below.

```

 $\tilde{x} = R \cdot \text{mid}(b(\mathbf{p}));$ 
 $D(\mathbf{p}) = \square\{R \cdot A(p) \mid p \in \mathbf{p}\};$ 
 $Z(\mathbf{p}) = \square\{R \cdot (b(p) - A(p)\tilde{x}) \mid p \in \mathbf{p}\};$ 
return  $\tilde{x} + \langle D(\mathbf{p}) \rangle^{-1} |Z(\mathbf{p})| \cdot [-1, 1];$ 

```

Algorithm 3. Skalna's Direct Method

Chevrons denote an interval extension of Ostrowski's comparison operator. It is defined by $\langle \mathbf{A} \rangle_{ii} = \langle \mathbf{A}_{ii} \rangle$, $\langle \mathbf{A} \rangle_{ij} = -|\mathbf{A}_{ij}|$ for $i \neq j$, where $\langle \mathbf{A}_{ii} \rangle$ is minimal absolute value, and $|\mathbf{A}_{ij}|$ is maximal absolute value [12]. It is required that $D(\mathbf{p})$ is an H-matrix which is equivalent to strong regularity of $A(\mathbf{p})$. If the condition is not fulfilled, the method produces unreliable results.

5 Numerical Examples

The results and computational times of generalized Rump’s fixed point iteration (GRM) and Skalna’s Direct method (SDM) are compared in this section. The overestimation measure $O_\omega = 100 \cdot (1 - w_{GRM}/w_{SDM})$ [13] is used to compare the results generated by those methods. Since the results of Rump’s method depend on the inflation parameter ε , different values of ε are considered: $\varepsilon = 0.1$ ([18]), $\varepsilon = 0.01$, and $\varepsilon = 1.0e^{-7}$.

Example 1 (Three-dimensional system)

$$\begin{pmatrix} -(p_1 + 1) * p_2 & p_1^2 * (p_3 - p_4) & -p_2 \\ p_5 / \sqrt{p_2 * p_4} & p_2 * (p_2 - p_3) & 1 \\ p_1 * p_2 & (p_1 - p_3) * p_5 & \sqrt{p_2} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_1 \\ p_1 \end{pmatrix} . \tag{5}$$

All parameters are considered to be uncertain with nominal values $p_1 = 1.2$, $p_2 = 2.2$, $p_3 = 0.51$, $p_4 = p_5 = 0.4$. Two cases of uncertainty are considered: 2% ($\pm 1\%$) and 16% ($\pm 8\%$). The results are presented in the following tables.

Table 1. 2% uncertainty; result obtained with standard (S) and new (N) multiplication formula

x	GRM _S	SDM _S	GRM _N	SDM _N
$\varepsilon = 0.1$				
x_1	[−4.39762, −3.858322]	[−4.39760, −3.858345]	[−4.39750, −3.858342]	[−4.39748, −3.858365]
x_2	[−1.581847, −1.360348]	[−1.581833, −1.360361]	[−1.581672, −1.360368]	[−1.581658, −1.36038]
x_3	[7.813244, 9.047074]	[7.813286, 9.047032]	[7.81328, 9.04683]	[7.81332, 9.046793]

Table 2. 16% uncertainty; standard (S) and new (N) multiplication formula

x	GRM _S	SDM _S	GRM _N	SDM _N
$\varepsilon = 0.1$				
x_1	[−17.65325, 9.38582]	[−16.78724, 8.51982]	[−17.44033, 9.17841]	[−16.55528, 8.29336]
x_2	[−8.83600, 5.90349]	[−8.35432, 5.42182]	[−8.67997, 5.75640]	[−8.18979, 5.26622]
x_3	[−17.88451, 34.77012]	[−16.27682, 33.16242]	[−17.48700, 34.36115]	[−15.84460, 32.71875]

Tables 3 and 4 show the percentage by which the GRM overestimates SDM, number of the GRM iterations, and the quotient of computational times.

Table 3. Example 1 (2%): comparison of the GRM and SDM results using O_ω measure, number of the GRM iterations, and the quotient of computational times

ε	O_ω	Iterations	time _{GRM} /time _{SDM}
0.1	0.26 – 0.38	2	1.1
0.01	0.02 – 0.03	3	1.5
0.0000001	0	5	1.9

Table 4. Example 1 (16%): comparison of the GRM and SDM results using O_ω measure, number of the GRM iterations, and the quotient of computational times

ε	O_ω	Iterations	$\text{time}_{GRM}/\text{time}_{SDM}$
0.1	6.34 – 6.79	6	1.3
0.01	0.52 – 0.56	11	1.9
0.0000001	0	17	3.0

The results of the GRM and SDM are very similar for small uncertainty and small ε , but the difference increases as the uncertainty and ε grows.

Example 2 (Simple planar frame). Consider a simple planar frame described in [8] (Section 4.2.1). Initially, the problem is solved with 1% uncertainty in all parameters. The results for moments and reactions of the planar frame system obtained using the SDM and the GRM methods are presented in Table 5.

Table 5. 1% uncertainty in all parameters

x	GRM	SDM
$\varepsilon = 0.1$		
x_1	[0.24464844, 0.25537031]	[0.24464851, 0.25537024]
x_2	[−0.51069763, −0.48933987]	[−0.51069754, −0.48933996]
x_3	[−1.01725791, −0.98281709]	[−1.01725785, −0.98281715]
x_4	[−0.76989609, −0.73016016]	[−0.76989590, −0.73016035]
x_5	[6.66820423, 6.83180202]	[6.66820453, 6.83180172]
x_6	[3.95928814, 4.04076186]	[3.95928829, 4.04076171]
x_7	[−0.68751300, −0.64632866]	[−0.68750909, −0.64633258]
x_8	[0.64632866, 0.68751300]	[0.64633258, 0.68750909]

For most components, the results of both methods coincide in their 6 leading digits. In this case, the overestimation of GRM with respect to SDM is almost negligible.

Table 6. 18% uncertainty in lengths and 6% uncertainty in load

x	GRM	SDM
$\varepsilon = 0.1$		
x_1	[0.13886897, 0.36720466]	[0.13936921, 0.36670443]
x_2	[−0.71730778, −0.29483949]	[−0.71674069, −0.29540658]
x_3	[−1.31437309, −0.70992145]	[−1.31397217, −0.71032236]
x_4	[−1.18360244, −0.33461846]	[−1.18226495, −0.33595596]
x_5	[5.41800409, 8.08401135]	[5.42004592, 8.08196951]
x_6	[3.36635653, 4.64984893]	[3.36730819, 4.64889727]
x_7	[−1.76378041, 0.32575072]	[−1.75510698, 0.31707729]
x_8	[−0.32575072, 1.76378041]	[−0.31707729, 1.75510698]

Table 6 reports the results obtained for the planar frame system with 18% uncertainty in lengths and 6% uncertainty in load. In this case, the results of both methods coincide only in their 2 leading digits for most solution components. The percentage by which GRM overestimates SDM, number of GRM iterations, and the quotient of computational times are given in Table 7.

Table 7. Example 2 (18%): comparison of GRM and SDM results using O_ω measure, number of GRM iterations, and the quotient of computational times

ε	O_ω	Iterations	$\text{time}_{GRM}/\text{time}_{SDM}$
0.1	0.13 – 0.83	2	1.1
0.01	0.01 – 0.09	3	1.3
0.0000001	0	8	2.2

6 Conclusions

Two methods for solving parametric interval linear systems were presented and compared in this study: the generalized Rump’s fixed point iteration and the single-step Skalna’s Direct Method. To show the performance of both methods, a couple of linear algebraic systems whose elements are functions of parameters belonging to given intervals were solved. The following characteristics were taken into account for the comparison purposes: accuracy of the approximation and computational time. It turned out from the numerical experiments (the overall conclusions are based on a much larger number of numerical experiments), that for small uncertainties and small ε -inflation, the accuracy of both methods were similar, but SDM performed faster. In fact, the results of the GRM converge to the results of SDM as ε tend to zero. When ε increases, the computational times become comparable, but SDM produces more accurate approximations, and the difference between approximations increases as uncertainty grows. Summarizing, the recommendation for practical applications is that SDM is better choice for accuracy and computational efficiency.

Acknowledgement. The author wishes to express her sincere thanks to all reviewers who dedicated their time and expertise to reviewing the manuscript and whose valuable remarks and suggestions have led to a substantial improvement of this paper.

References

- 1. Akhmerov, R.R.: Interval-affine Gaussian algorithm for constrained systems. *Reliable Computing* 11(5), 323–341 (2005)
- 2. El-Owny, H.: Parametric Linear System of Equations, whose Elements are Non-linear Functions. In: 12th GAMM - IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics, vol. 16 (2006)

3. Garloff, J., Popova, E.D., Smith, A.P.: Solving Linear Systems with Polynomial Parameter Dependency in the Reliable Analysis of Structural Frames. To appear in Proceedings of the 2nd International Conference on Uncertainty in Structural Dynamics, Sheffield, UK, June 15-17 (2009)
4. Vu, X.-H., Sam-Haroud, D., Faltings, B.: A Generic Scheme for Combining Multiple Inclusion Representations in Numerical Constraint Propagation. Technical Report No. IC/2004/39, Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland (2004)
5. Kolev, L.V.: Solving Linear Systems whose Elements are Non-linear Functions of Intervals. *Numerical Algorithms* 37, 213–224 (2004)
6. Kolev, L.V.: A new method for global solution of systems of non-linear equations. *Reliable Computing* 4, 125–146 (1998)
7. Kolev, L.V.: Automatic computation of a linear interval enclosure. *Reliable Computing* 7, 17–18 (2001)
8. Kulpa, Z., Pownuk, A., Skalna, I.: Analysis of linear mechanical structures with uncertainties by means of interval methods. *Computer Assisted Mechanics and Engineering Sciences* 5(4), 443–477 (1998), <http://andrzej.pownuk.com/publications/IntervalEquations.pdf>
9. Messine, F.: Extensions of Affine Arithmetic: Application to Unconstrained Global Optimization. *Journal of Universal Computer Science* 8(11), 992–1015 (2002)
10. Miyajima, S., Miyata, T., Kashiwagi, M.: On the Best Multiplication of the Affine Arithmetic. *Transactions of the Institute of Electronics, Information and Communication Engineers* J86-A(2), 150–159 (2003)
11. Muhanna, R.L., Zhang, H., Mullen, R.L.: Interval Finite Elements as a Basis for Generalized Models of Uncertainty in Engineering Mechanics. *Reliable Computing* 13(2), 173–194 (2007)
12. Neumaier, A.: *Interval Methods for Systems of Equations*, pp. xvi–255. Cambridge University Press, Cambridge (1990)
13. Popova, E.D.: On the Solution of Parametrised Linear Systems. In: Kraemer, W., von Wolff Gudenberg, J. (eds.) *Scientific Computing, Validated Numerics, Interval Methods*, pp. 127–138. Kluwer Acad. Publishers, Dordrecht (2001)
14. Popova, E.: Generalizing the Parametric Fixed-Point Iteration. *Proceedings in Applied Mathematics & Mechanics (PAMM)* 4(1), 680–681 (2004)
15. Rohn, J., Rex, G.: Enclosing solutions of linear equations. *SIAM Journal Numerical Analysis* 35(2), 524–529 (1998)
16. Rump, S.M.: New Results on Verified Inclusions. In: Miranker, W.L., Toupin, R.A. (eds.) *Accurate Scientific Computations. LNCS*, vol. 235, pp. 31–69. Springer, Heidelberg (1986)
17. Rump, S.M.: Verification methods for dense and sparse systems of equations. In: Herzberger, J. (ed.) *Topics in Validated Computations*, pp. 63–135. North-Holland, Amsterdam (1994)
18. Rump, S.M.: A note on epsilon-inflation. *Reliable Computing* 4, 371–375 (1998)
19. Shary, S.P.: Solving tied interval linear systems. *Sibirskii Zhurnal Vychislitel'noi Matematiki* 7(4), 363–376 (2004)
20. Skalna, I.: A Method for Outer Interval Solution of Systems of Linear Equations Depending Linearly on Interval Parameters. *Reliable Computing* 12(2), 107–120 (2006)
21. Skalna, I.: Direct method for solving parametric interval linear systems with non-affine dependencies. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) *PPAM 2009. LNCS*, vol. 6068, pp. 485–494. Springer, Heidelberg (2010)

Numerical Investigation of the Upper Bounds on the Convective Heat Transport in a Heated from below Rotating Fluid Layer

Nikolay Vitanov

Institute of Mechanics, Bulgarian Academy of Sciences, akad. G. Bonchev Str., Bl. 4,
1113 Sofia, Bulgaria

vitanov@imbm.bas.bg

<http://www.imbm.bas.bg/index.php/en-us/vitanov>

Abstract. We apply the Galerkin method in order to obtain numerical solution of the Euler - Lagrange equations for the variational problem for the upper bounds on the convective heat transport in a fluid layer under the action of intermediate and strong rotation. The role of the numerical investigation in such kind of variational problems is to obtain the upper bounds for the case of small and intermediate values of the Rayleigh and Taylor numbers in addition to the analytical asymptotic theory which leads to the upper bounds for the case of large values of the above two characteristic dimensionless numbers. The application of the Galerkin method reduces the Euler - Lagrange equations to a system of nonlinear algebraic equations. This system is solved numerically by the Powel hybrid method. We observe that the Powel hybrid method guarantees satisfactory fast rate of convergence from the guess solution to the solution of the system of equations. We present and discuss several results from the numerical computations.

1 Introduction

The rotation and the thermal convection are important for the fluid motions in the planetary atmospheres and Earth's oceans. Thus it is of interest to study the thermal convection in a rotating fluid layer [1]. Here we present a numerical investigation of such a system based on the optimum theory of turbulence.

The optimum theory of turbulence [2] - [5] leads to rigorous numerical or analytical estimates of the upper bounds on the turbulent quantities directly from the nonlinear Navier-Stokes equations. Such results are obtained on the basis of variational problem obtained by means of finite number of moment equations (power integrals) which are derived from the Navier-Stokes equations. In such a way the energy balance of the real flow is retained and the solutions of the Euler - Lagrange equations of the variational problem lead to upper bounds on characteristic quantities for the turbulent flows [6], [7].

Below we shall apply the Howard-Busse method of the optimum theory of turbulence [2], [6]. This method was applied to many cases of fluid flows and thermal convection [8] - [12]. We shall investigate the convective heat transport

in the rotating about a vertical axis horizontal layer of fluid for the case of intermediate and large rotation rates i.e. for such values of the Taylor number for which the rotation begins to influence the internal layers of the fields which are solutions of the Euler-Lagrange equations of the corresponding variational problem. Thermal convection at large Rayleigh numbers is strongly nonlinear. No exact analytical description of its is known and thus the numerical results are extremely useful [13] - [17].

2 Mathematical Formulation of the Problem

We consider a heated from below horizontal layer of fluid. The layer rotates about the vertical axis with a constant angular velocity Ω . The layer sizes are infinite in horizontal directions. Thermal processes are modeled by the Boussinesq approximation to the equations of the fluid flow [1]. Denoting the layer thickness as d , the thermometric conductivity and kinematic viscosity of the fluid as κ and ν , the acceleration of the gravity as g , the temperature difference between the upper and lower fluid boundary as ΔT and the density of the fluid as ρ , and taking d as an unit for length, κ/d as unit for velocity, d^2/κ as unit for time and $\rho\nu\kappa/d^2$ as unit for pressure we can write the dimensionless form of the Boussinesq equations

$$\frac{1}{P} \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\frac{\sqrt{Ta}}{2} \nabla p + \nabla^2 \mathbf{u} + R T \mathbf{k} + \sqrt{Ta} \mathbf{u} \times \mathbf{k} \quad (1)$$

$$\frac{\partial \Theta}{\partial t} + \mathbf{u} \cdot \nabla \Theta = \nabla^2 \Theta; \quad \nabla \cdot \mathbf{u} = 0 \quad (2)$$

The boundary conditions are stress-free at $\pm z = 1/2$: $u_3 = \partial^2 u_3 / \partial z^2 = T = 0$. $P = \nu/\kappa$ is the Prandtl number, $Ta = (2\Omega d^2/\nu)^2$ is the Taylor number $R = (\gamma g \Delta T d^3)/(\kappa \nu)$ is the Rayleigh number, γ is the coefficient of thermal expansion, p is the pressure, and $\mathbf{k}$ is the unit vector in the direction opposite to the gravity. The quantity Θ in (2) is the total temperature field and T is the deviation of the temperature field from its horizontal mean $\Theta = \bar{\Theta} + T$. Here and below we shall use averages of the quantities over the planes $z = \text{const}$ (denoted as $\bar{q}$) and over the fluid layer (denoted as $\langle q \rangle$). A variational problem can be formulated by two moment equations obtained on the basis of the Boussinesq equations [12]. We shall assume that all necessary horizontal averages of the functions, describing the flow exist, that the horizontal averages of the fluctuation quantities vanish, and that the flow is statistically steady in time and homogeneous in the horizontal averages. Our goal is to obtain an upper bound on the convective heat transport through the fluid layer i.e. on the Nusselt number

$$Nu = 1 + \frac{\langle u_3 T \rangle}{R} \quad (3)$$

The two moment equations are

$$\langle |\nabla \mathbf{u}|^2 \rangle = R \langle u_3 T \rangle; \quad \langle |\nabla T|^2 \rangle = \langle u_3 T \rangle^2 - \overline{\langle u_3 T^2 \rangle} + \langle u_3 T \rangle \quad (4)$$

When Prandtl number is infinite the Navier-Stokes equation becomes linear and we can include it as a constraint in the variational problem. We shall take into account the equation of continuity by the general representation of a solenoidal field $\mathbf{u}$ in terms of a poloidal and a toroidal component $\mathbf{u} = \nabla \times (\nabla \times \mathbf{k}\phi) + \nabla \times \mathbf{k}\psi$. Let us perform the rescaling $\mathbf{u} = \mu^{1/2} \langle w\theta \rangle^{-1/2} \mathbf{v}$; $T = \mu^{1/2} \langle w\theta \rangle^{-1/2} R^{-1} \theta$, where the z -component of the rescaled velocity field $\mathbf{v}$ is denoted as w . We obtain from (4)

$$R = \frac{\langle |\nabla\theta|^2 \rangle}{\langle w\theta \rangle} + \mu \frac{\langle (\overline{w\theta} - \langle w\theta \rangle)^2 \rangle}{\langle w\theta \rangle^2} \quad (5)$$

Let us introduce the toroidal-poloidal decomposition into the Navier-Stokes equation ($P = \infty$). Taking the z -component of the horizontal curl and z -component of the double curl of the result we obtain the relationships

$$\nabla^2 f + \sqrt{Ta} \frac{\partial w}{\partial z} = 0; \quad \nabla^4 w + \nabla_1^2 \theta - \sqrt{Ta} \frac{\partial f}{\partial z} = 0 \quad (6)$$

where $f = -\nabla_1 \psi$ is the vertical component of the vorticity. By means of Lagrange multipliers p^* and q^* we include (6) in the variational problem which can be formulated as follows

Find the minimum $R(\mu)$ of the variational functional

$$\begin{aligned} \mathcal{F} = \frac{\langle |\nabla\theta|^2 \rangle}{\langle w\theta \rangle} + \mu \frac{\langle (\overline{w\theta} - \langle w\theta \rangle)^2 \rangle}{\langle w\theta \rangle^2} - \left\langle p^* \left(\nabla^2 f + \sqrt{Ta} \frac{\partial w}{\partial z} \right) \right\rangle - \\ - \left\langle q^* \left(\nabla^4 w + \nabla_1^2 \theta - \sqrt{Ta} \frac{\partial f}{\partial z} \right) \right\rangle = 0 \end{aligned} \quad (7)$$

among all fields w, θ, f that satisfy the boundary conditions

$$w = \theta = \frac{\partial^2 w}{\partial z^2} = \frac{\partial f}{\partial z} = 0 \quad (8)$$

at $z = \pm 1/2$.

The Euler-Lagrange equations for the above functional are

$$\langle |\nabla\theta|^2 \rangle \theta + 2\theta[\mu(\overline{w\theta} - \langle w\theta \rangle) - R\langle w\theta \rangle] - \langle w\theta \rangle^2 \left(\nabla^4 q^* - \sqrt{Ta} \frac{\partial p^*}{\partial z} \right) = 0 \quad (9)$$

$$\langle |\theta|^2 \rangle w - 2\langle w\theta \rangle \nabla^2 \theta + 2w[\mu(\overline{w\theta} - \langle w\theta \rangle) - R\langle w\theta \rangle] - \langle w\theta \rangle^2 \nabla_1^2 q^* = 0 \quad (10)$$

$$\nabla^2 p^* + \sqrt{Ta} \frac{\partial q^*}{\partial z} = 0; \quad \nabla^4 w + \nabla_1^2 \theta = \sqrt{Ta} \frac{\partial f}{\partial z} \quad (11)$$

$$\nabla^2 f + \sqrt{Ta} \frac{\partial w}{\partial z} = 0 \quad (12)$$

3 Details on the Numerical Solution

We eliminate the Lagrange multipliers and introduce the $1 - \alpha$ - solutions of the variational problem: $w = w_1(z)\phi(x, y)$; $\theta = \theta_1(z)\phi(x, y)$; $f = f_1(z)\phi(x, y)$. where $\nabla_1^2 \phi = -\alpha_1^2 \phi$. Due to the homogeneity of the Euler-Lagrange equations we can impose the requirement $\mu = \langle w_1 \theta_1 \rangle$. On the basis of our experience from the numerical investigation of the case of finite Prandtl number fluid for w_1 , θ_1 and f_1 we use the following symmetric representations: $w_1(z) = \sum_{m=1}^M a_m \sin[2(m-1)\pi(z+1/2)]$; $\theta_1(z) = \sum_{m=1}^M b_m \sin[(2m-1)\pi(z+1/2)]$; $f_1(z) = \sum_{m=1}^M c_m \cos[(2m-1)\pi(z+1/2)]$, where the parameter M has to be truncated in such a way that the solutions do not depend in any significant way on it. The largest value of M used in calculations was $M = 180$ and we adopted the criterion that M is sufficiently large if the Nusselt number Nu changes by less than 0.1% when M is replaced by $M - 5$.

The result of the theoretical considerations so far is a reduction of the system of nonlinear integro-differential Euler-Lagrange equations to a system of nonlinear algebraic equations with (i) no preexisting knowledge about the solution; (ii) no simple way to suggest a starting vector (a guess of the solution); and (iii) no simple way to reduce the search area. In order to obtain good enough approximation of the solution of the Euler - Lagrange equations the number M of the modes (connected to the size of the nonlinear algebraic system) must be large enough. When Taylor number is fixed M increases rapidly with increasing Rayleigh number. When Rayleigh number is fixed and Taylor number increases then the rotation inhibits the thermal convection and M can decrease.

For the problem we solve it is very important that the chosen numerical method of solution is robust and especially it is important that the method is capable to find a solution even if the starting vector is far from the vector of the solution. In other words the method must have a good convergence. Our experience has shown that for the class of nonlinear systems connected to the problem of upper bounds on the turbulent thermal convection the Powel hybrid method [18,19] satisfies the necessary conditions in the most convenient manner. The nonlinear systems connected to the variational problems for thermal convection always possess a zero solution which corresponds to the state of thermal conduction and gives a lower bound on the convective heat transport. The Powel hybrid method successfully avoids this solution and converges easily to the class of the non-zero solutions of the variational problem (known as multi-wavenumber solutions) which describe the different regimes of the thermal convection. Another important property of the Powel hybrid method is that it converges satisfactory fast even for large number of equation of the solved system of nonlinear algebraic equations.

A brief illustration of important concepts of the numerical method is as follows. We solve the system of nonlinear algebraic equations $\mathbf{f}(\mathbf{x}) = 0$. It is well known that in the Gauss-Newton method one makes a linear approximation of $\mathbf{f}$ in the neighborhood of $\mathbf{x}$ as follows:

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) \simeq \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{h} \quad (13)$$

At the starting point of the Powel method the Jacobian J is approximated by finite differences. Then the Jacobian is updated by the rank-1 method of Broyden. Applied to (13) the update is as follows:

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) \simeq \mathbf{f}(\mathbf{x}) + B(\mathbf{x})\mathbf{h} \quad (14)$$

where $B(\mathbf{x})$ is the current approximation of the Jacobian $J(\mathbf{x})$. For the next iteration step we calculate B_{new} such that

$$\mathbf{f}(\mathbf{x}_{new} + \mathbf{h}) \simeq \mathbf{f}(\mathbf{x}_{new}) + B_{new}\mathbf{h} \quad (15)$$

We request that (15) hold with equality for $\mathbf{h} = \mathbf{x} - \mathbf{x}_{new}$. Then the broyden rank-1 update is

$$B_{new} = B + \left\{ \frac{\mathbf{f}(\mathbf{x}_{new}) - \mathbf{f}(\mathbf{x}) - B\mathbf{h}}{(\mathbf{x}_{new} - \mathbf{x})^T(\mathbf{x}_{new} - \mathbf{x})} \right\} (\mathbf{x}_{new} - \mathbf{x})^T \quad (16)$$

In addition the correction at each step is given as combination of the Gauss-Newton and steepest descent direction. In practice this leads to very good convergence for the class of problems we have to solve.

4 Results and Discussion

The Galerkin method and the Powell hybrid method are a very good combination of methods for numerical investigation of the variational problems for obtaining upper bounds on the convective heat transport in fluid layers under the action of rotation. Fig. 1 illustrates this finding. As we can see the number M of the components needed for satisfactory description of the profiles of the optimum fields decreases with increasing Taylor number when the Rayleigh number is fixed. Thus the more rapid rotation requires less numerical efforts. The reason for this is that the rapid rotation inhibits the thermal convection. This results in thicker boundary layer and because of this we need smaller number of components M for description of the profiles of these layers. The situation is opposite to the case without rotation. There the rotation does not inhibit the convection and the number M increases steadily according to a power law with increasing Rayleigh number. The number of components M for the case with presence of rotation does not follow a power law as it can be seen from Fig. 1.

Several results for the numerical calculation of the profiles of the optimum fields (the fields which are solutions of the Euler - Lagrange equations of the variational problem) are shown in Fig. 2. We note that the inhibiting effects of the rotation is expressed by the lack of peaks of the profile of the field w_1 and in the more slow development of the peaks in the field θ_1 . Because of the fact that the Galerkin method and Powell hybrid method are convenient tools for numerical investigation of variational problems for the thermal convection in presence of rotation we obtain smooth profiles for the optimum fields without much computational efforts.

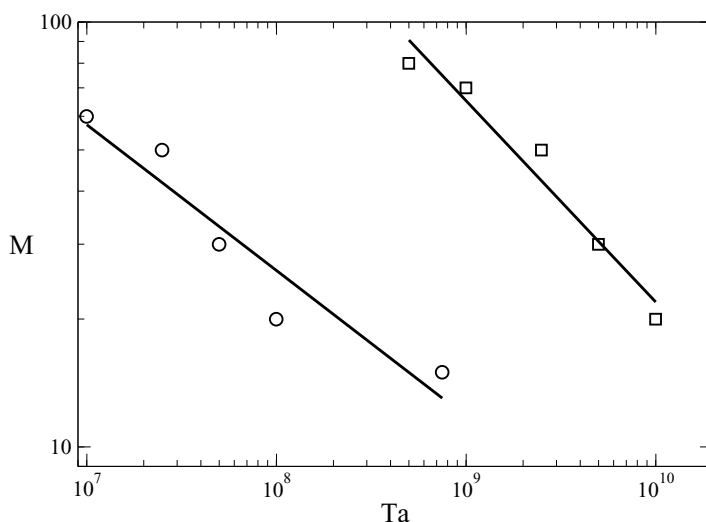


Fig. 1. Needed number of components M for the numerical investigation of the upper bounds on the convective heat transport. M is presented as function of the Taylor number Ta for two fixed values of the Rayleigh number R . Circles: $R = 10^7$. Squares: $R = 10^8$. For orientation two power laws are presented with lines. Power law connected to the data shown with circles: $M = 14.4 \cdot 10^3 \cdot Ta^{-0.343}$. Power law connected to the data shown by squares: $M = 1.2 \cdot 10^6 \cdot Ta^{-0.474}$.

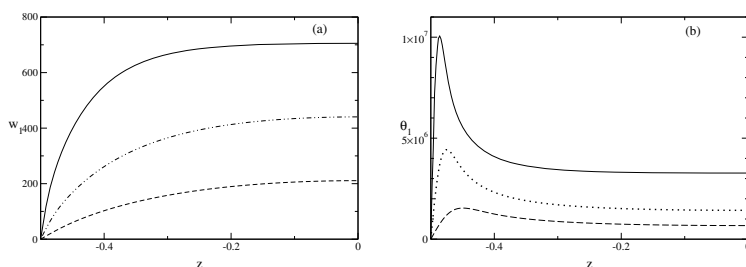


Fig. 2. Selected profiles of the optimum fields. Because of the symmetry of the profiles only the region between $z = -0.5$ and $z = 0$ is shown. Figure (a): Influence of Rayleigh number on the optimum field w_1 . $Ta = 10^9$. From bottom to the top: dashed line : $R = 2.5 \cdot 10^7$; dot-dashed line: $R = 5 \cdot 10^7$; solid line: $R = 10^8$. We observe that the strong rotation leads to slowing the development of the field w_1 . Figure (b): Influence of Rayleigh number on the optimum field θ_1 . All values of the Rayleigh and Taylor numbers are as in the Figure (a).

Finally the used numerical methodology allows us to determine with very good accuracy the thickness δ of the boundary layers of the optimum fields and to investigate the changes in δ with changing values of the Rayleigh and Taylor numbers. In such a way we obtain power-law dependence of the thickness of

the boundary layers from R and Ta . For an example we have obtained that the thickness of the boundary layer of the field θ_1 follows the law $\delta = 6.28 \cdot 10^{-7} Ta^{0.48}$ when the Rayleigh number is fixed at $R = 10^8$. When $R = 10^9$ the corresponding power-law for the same field is $\delta = 1.29 \cdot 10^{-10} Ta^{0.75}$.

5 Concluding Remarks

In this paper we have shown that the combination of Galerkin method and the Powell hybrid method is very appropriate for investigation of the variational problems of the optimum theory of turbulent and non-turbulent thermal convection. The Galerkin method leads to a reduction of the nonlinear integro-differential Euler - Lagrange equations of the variational problem to a system of nonlinear algebraic equations. This system can be solved with the help of the Powell hybrid method which ensures very good convergence for the class of systems of algebraic equations connected to the variational problems of the thermal convection. The presence of rotation makes the combination from Galerkin and Powell hybrid method even more appropriate as the increasing rotation leads to decrease of the number of nonlinear algebraic equations we have to solve. In addition we are able to determine precise the profiles of the optimum fields and to extract directly the power laws which govern the evolution of the thickness of the boundary layers of the optimum fields when the Rayleigh and Taylor numbers change their values.

Finally the numerical investigation is very useful addition to the analytical asymptotic theory for the case of non-asymptotic values of the Rayleigh and Taylor numbers. If we continue the asymptotic Eq. (51b) from [16] back to the non-asymptotic values $R = 10^9$, $Ta = 10^{11}$ we shall obtain for the upper bound on the heat transport $Nu^* = 131.78$. The numerical solution of the Euler - Lagrange equations leads to $Nu = 13.3$. This is an illustration of the fact that for non-asymptotic values of the Rayleigh and Taylor numbers the numerical results lead to lower upper bound in comparison to the analytical ones. When the Rayleigh and Taylor numbers tend to the asymptotic large values the numerical upper bounds approach from below the analytical upper bounds as for an example for the case discussed in [12].

Acknowledgment

This research was supported by the Grant DO 02/338 - 22.12.2008 of the National Fund for Scientific Researches of Republic of Bulgaria.

References

1. Chandrasekhar, S.: Hydrodynamics and Hydromagnetic Stability. Dover, New York (1981)
2. Howard, L.N.: Heat transport by turbulent convection. J. Fluid Mech. 17, 405–432 (1963)

3. Hoffmann, N.P., Vitanov, N.K.: Bounds on energy dissipation in turbulent shear flow under the action of rotation. *Phys. Lett. A* 255, 277–286 (1999)
4. Vitanov, N.K.: Upper bounds on the heat transport in a porous layer. *Physica D* 136, 322–339 (2000)
5. Vitanov, N.K., Busse, F.H.: Upper bound on the heat transport in a heated from below fluid layer. *Springer Proceedings in Physics* 101, 37–40 (2005)
6. Busse, F.H.: On Howard's upper bound for heat transport by turbulent convection. *J. Fluid. Mech.* 37, 457–477 (1969)
7. Vitanov, N.K., Busse, F.H.: Bounds on heat transport in a horizontal fluid layer with stress-free boundaries. *Zeitschrift fur Angewandte Mathematik und Physik (ZAMP)* 48, 310–324 (1997)
8. Straus, J.M.: On the upper bounding approach to thermal convection at moderate Rayleigh numbers. II. Rigid boundaries. *Dyn. Atm. Oceans* 1, 77–90 (1976)
9. Vitanov, N.K.: Upper bound on the heat transport in a horizontal layer of infinite Prandtl number. *Phys. Lett. A* 248, 338–346 (1998)
10. Vitanov, N.K.: Upper bound on the heat transport in a layer of fluid of infinite Prandtl number, rigid lower boundary and stress-free upper boundary. *Phys. Rev. E* 61, 956–959 (2000)
11. Vitanov, N.K.: Convective heat transport in a fluid layer of infinite Prandtl number: Upper bounds for the case of rigid lower boundary and stress-free upper boundary. *European Physical Journal B* 15, 349–355 (2000)
12. Vitanov, N.K.: Numerical upper bounds on convective heat transport in a layer of fluid of finite Prandtl number. Confirmation of Howard's analytical asymptotic single-wave-number bound. *Physics of Fluids* 17, Article Number 105106 (2005)
13. Vitanov, N.K., Busse, F.H.: Bounds on the convective heat transport in a rotating layer. *Phys. Rev. E* 63, Article Number 016303 (2001)
14. Vitanov, N.K.: Convective heat transport in a rotating layer of infinite Prandtl number: Optimum fields and upper bounds on Nusselt number. *Phys. Rev. E* 67, Article Number 026322 (2003)
15. Vitanov, N.K.: Upper bounds on convective heat transport in a rotating fluid layer of infinite Prandtl number: Case of intermediate Taylor numbers. *Phys. Rev. E* 62, 3581–3591 (2000)
16. Vitanov, N.K.: Upper bounds on convective heat transport in a rotating fluid layer of infinite Prandtl number: Case of large Taylor numbers. *European Physical Journal B* 23, 249–266 (2001)
17. Vitanov, N.K.: Optimum fields and upper bounds for nonlinear convection in rapidly rotating fluid layer. *European Physical Journal B* 73, 265–273 (2010)
18. Powell, M.J.D.: A hybrid method for nonlinear algebraic equations. In: Rabinowitz, P. (ed.) *Numerical Methods for Nonlinear Algebraic Equations*, pp. 87–114. Gordon and Breach, New York (1970)
19. Madsen, K., Nielsen, H.B., Tingeloff, O.: *Methods for non-linear least squares problems. Informatics and Mathematical Modeling*. Technical University of Denmark (2004)

Author Index

- Angelova, Maria 224
Asenov, Asen 41
Atanasova, P.Kh. 347
Atanasov, Krassimir 232, 240, 248, 256
Atanassova, Liliya 232
Atanassova, Vassia 240
Ayuso, B. 353
- Bogachev, Andrey 215
Bouvry, Pascal 297
Boyadjiev, Todor L. 347, 361
Bradji, Abadallah 369
Březina, Jan 125, 420
- Česenek, Jan 1
Chernogorova, Tatiana 377
Christov, Christo I. 386
Csendesí, Ádam 77
- Dimov, Ivan 50, 95, 198
Dimova, Milena 395
Dimova, Stefka 395, 428
Dinis, M.L. 60
Dobrinkova, Nina 133
Duda, Jerzy 305
- Ebel, Adolf 174
Elbern, Hendrik 174
Elkin, N.N. 404
Etropolska, Iglíka 141
- Faragó, István 198
Feistauer, Miloslav 1
Fidanova, Stefka 248, 256
Fiúza, António 60, 190
Fríncu, Marc 321
Fuhrmann, Jürgen 369
Fujimoto, Noriyuki 264
- Gadzhev, Georgi 150
Ganev, Kostadin 141, 150, 215
Georgiev, Ivan 353, 412
Georgiev, Krassimir 158, 198
- Georgieva, Rayna 50
Gómez-Pulido, Juan A. 313
Goodnick, S.M. 103, 118
- Hatzigeorgiou, George 477
Havasi, Ágnes 198
Hokr, Milan 125, 420
Hossain, A. 118
Hristov, Ivan 428
Hristov, Vladimir 437
- Iliev, Anton 437
Iliev, Oleg 329, 338
Ishimura, Naoyuki 445
- Jordanov, Georgi 133, 150
- Kandilarov, Juri D. 453
Kochev, Nikolay T. 182
Koleva, Miglena N. 445, 461
Kolkovska, Natalia T. 386, 469
Kopal, Jiří 420
Kraus, J. 353
Kumar, Sunil 486
Kyurkchiev, Nikolay 437
- Latz, Arnulf 329
Liolios, Angelos 477
Liolios, Asterios 477
Liolios, Konstantinos 167
Lirkov, Ivan 68
Lymbery, Maria 412
- Magdics, Milan 77
Makarov, Alexander 87
Mandel, Jan 133
Margenov, Svetozar 338, 412
Marinov, Pencho 248, 256
Melemov, Hristo T. 361
Miloshev, Nikolai 141, 150, 215
- Napartovich, A.P. 404
Nedjalkov, M. 95
- Ostromsky, Tzvetan 198

- Panetsos, Panagiotis 477
 Pedroso, João Pedro 272
 Pencheva, Tania 224
 Penev, Kalin 280
 Penzov, Anton 77
 Popov, P. 338
 Prodanova, Maria 141, 150, 215

 Radev, Stefan 167, 477
 Rálek, Petr 420
 Raleva, K. 103, 118
 Rao, S. Chandra Sekhara 486
 Resteanu, Cornel 207
 Roeva, Olympia 289
 Ruzhekov, Anton 280

 Sabelfeld, Karl 14
 Sánchez-Pérez, Juan M. 313
 Selberherr, Siegfried 87, 95
 Seredynski, Marcin 297
 Shukrinov, Yu.M. 347
 Skalna, Iwona 305, 494
 Slavov, Kiril 141, 215
 Slavov, Tsonyo 289
 Soeiro de Carvalho, J.M. 190
 Spiridonov, Valery 215
 Stefanov, Stefan K. 110
 Stoilova, Stanislava 68
 Strunk, Achim 174
 Sverdlov, Viktor 87

 Syrakov, Dimiter 141, 150, 215
 Szirmay-Kalos, Laszlo 77

 Terziyski, Atanas T. 182
 Torrecilla-Pinero, Fernando 313
 Torrecilla-Pinero, Jesús A. 313
 Tóth, Balazs 77
 Trandafir, Romica 207
 Tsihrintzis, Vassilios 167
 Tsutsui, Shigeyoshi 264
 Tzonkov, Stoyan 224

 Vabishchevich, Petr N. 29
 Valkov, Radoslav L. 377, 453
 Vasileska, D. 103, 118
 Vasileva, Daniela 386
 Vega-Rodríguez, Miguel A. 313
 Vila, Maria Cristina 190
 Vitanov, Nikolay 502
 Vulkov, Lubin G. 445, 461
 Vutov, Y. 338
 Vysotsky, D.V. 404

 Zaharie, Daniela 321
 Zamfirache, Flavia 321
 Zausch, Jochen 329
 Zemlyanaya, E.V. 347
 Zikatanov, L. 353
 Zlatev, Zahari 158, 198