Springer Proceedings in Mathematics & Statistics

Monica G. Cojocaru
Ilias S. Kotsireas
Roman N. Makarov
Roderick V. N. Melnik
Hasan Shodiev *Editors*

Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science

Springer

Springer Proceedings in Mathematics & Statistics

Volume 117

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Monica G. Cojocaru • Ilias S. Kotsireas
Roman N. Makarov • Roderick V. N. Melnik
Hasan Shodiev
Editors

Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science

Springer

Editors

Monica G. Cojocaru
Department of Mathematics
& Statistics
University of Guelph
Guelph, ON, Canada

Ilias S. Kotsireas
Department of Physics
& Computer Science
Wilfrid Laurier University
Waterloo, ON, Canada

Roman N. Makarov
Department of Mathematics
and MS2Discovery Interdisciplinary
Research Institute
Wilfrid Laurier University
Waterloo, ON, Canada

Roderick V. N. Melnik
Department of Mathematics
& MS2Discovery Interdisciplinary
Research Institute
Wilfrid Laurier University
Waterloo, ON, Canada

Hasan Shodiev
MS2Discovery Interdisciplinary
Research Institute
Wilfrid Laurier University
Waterloo, ON, Canada

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-12306-6 ISBN 978-3-319-12307-3 (eBook)
DOI 10.1007/978-3-319-12307-3

Library of Congress Control Number: 2015936922

Mathematics Subject Classification: 34-XX, 92BXX, 37-XX, 37-XX, 00A79, 91AXX, 91AXX, 34-XX, 00A69, 34-XX, 76-XX, 92C05, 92BXX, 91GXX, 93-XX, 91AXX, 62-XX, 92-XX, 65-XX, 35LXX, 92D25, 00A69, 35-XX, 62-XX, 00A69, 62-XX, 00A69, 00A69, 34-XX, 34-XX, 90-XX (Optimization) 97-XX Combinatorics, 91GXX, 35LXX, 92BXX, 00A69, 97-XX, 00A69, 34-XX, 74-XX, 65-XX, 65-XX, 65-XX, 65-XX, 65-XX, 60-XX, 65-XX

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

Applied Mathematics, Modelling and Computational Science (AMMCS)-2013 was an interdisciplinary international conference, the second in a series of AMMCS meetings held in Waterloo, Ontario, Canada and hosted at Wilfrid Laurier University's Waterloo campus. The series aims at promoting interdisciplinary research and collaboration involving mathematical and computational sciences, and highlighting recent advances in AMMCS. AMMCS-2013 was held on August 26–30, 2013 and was organized in cooperation with AIMS and SIAM, with support from the Fields Institute in Toronto. The AMMCS-2013 book of abstracts can be found online at http://www.ammcs2013.wlu.ca/ and on behalf of all conference organizers and participants we wish to extend a heartfelt thanks to Mr. Dalibor D. Dvorski who devoted an enormous amount of time to edit it.

The first conference in the AMMCS series was held on July 25–29, 2011 and hosted at Wilfrid Laurier University's Waterloo campus. The AMMCS-2011 book of abstracts can be found online at http://www.ammcs2011.wlu.ca/ and on behalf of all conference organizers and participants we wish to extend a heartfelt thanks to Mr. Cameron Davidson-Pilon who devoted an enormous amount of time to edit it. The proceedings of AMMCS-2011 have been published in *Advances in Mathematical and Computational Methods: Addressing Modern Challenges of Science, Technology, and Society,* Ilias Kotsireas, Roderick Melnik, Brian West (Editors), AIP Conference Proceeding 1368, (2011). Selected papers received from the AMMCS-2011 plenary speakers have been published in *Advances in Applied Mathematics, Modelling, and Computational Science*, edited by Roderick Melnik and Ilias Kotsireas (Fields Institute Communications, 66, Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, 2013).

The AMMCS series of conferences is meant to provide a platform of interaction between applied scientists at all levels in various fields and aims to keep the very interdisciplinary nature of its content and sessions as its distinctive mark. There were many young scientists at AMMCS-2013, both as presenters and organizers. There was a poster session with specific prizes to encourage young trainees in the modelling field towards honing their communication skills on topics from their areas. AMMCS-2013 also managed to successfully pair this infusion of young researchers with an extensive list of prestigious plenary and semi-plenary speakers, whose presence added an enthusiastic layer of interaction between all participants.

AMMCS-2013 featured 42 special sessions and 10 contributed sessions with a total of 650 participants from 40 countries. The program of the conference was rich and varied with over 550 talks and posters being presented. Highlights were the invited plenary talks given by Peter Carr (Morgan Stanley), Emily Carter (Princeton University), Ronald Coifman (Yale University), Marty Golubitsky (Ohio State University), Vaughan Jones (Vanderbilt University), Lila Kari (Western University), Dimitris Giannakis and Andrew Majda (New York University), George Papanicolau (Stanford University), Panos Pardalos (University of Florida), Michael Sigal (University of Toronto), and Godfried Toussaint (NYUAD/M.I.T./McGill).

The present proceedings contains 78 refereed papers that were submitted to the AMMCS-2013 Editorial Team after the conference. The papers in this volume were carefully screened and we are grateful to a number of researchers around the world that acted as referees, offering pertinent referee reports to the authors of submitted papers.

We thank all the people who participated in the AMMCS-2013 conference and presented excellent talks, as well as all who contributed to the organization of the conference. We appreciate the help of students during the conference. We gratefully acknowledge the financial support for the conference by the Fields Institute and the Office of Research Services of Wilfrid Laurier University. Finally, we want to express our gratitude to Springer-Verlag for publishing this volume.

Waterloo, Ontario, Canada
July 2014

Monica G. Cojocaru
Ilias S. Kotsireas
Roman N. Makarov
Roderick V. N. Melnik
Hasan Shodiev

Contents

Contributors

Radek Štefan Department of Concrete and Masonry Structures, Faculty of Civil Engineering, Czech Technical University in Prague, Prague 6, Czech Republic

Abdullahi Rashid Adem International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mmabatho, South Africa

Hani Ali Institut Jean le Rond d'Alembert, CNRS & UPMC Université Paris 06, Paris, France

D. Almaatani Department of Mathematics and Computer Sciences, Laurentian University, Sudbury, Canada

A. Almowanes Laurentian University, Sudbury, ON, Canada

M. S. Alwan University of Waterloo, Waterloo, ON, Canada

Michael Andrews University of Guelph, Guelph, ON, Canada

M. Baboulin Inria and University of Paris-Sud, Orsay, France

A. Bahramian Department of Chemical Engineering, University of Tehran, Tehran, Iran

A. D. Bandrauk Laboratoire de chimie théorique, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, QC, Canada

K. Beauvais Nonlinear Dynamical Systems Group, Department of Mathematics, San Diego State University, San Diego, CA, USA

Aziz Belmiloudi IRMAR-INSA of Rennes, Rennes, France

Michal Beneš Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Prague 6, Czech Republic

M. I. Berenguer Department of Applied Mathematics, University of Granada, Granada, Spain

Faina S. Berezovskaya Department of Mathematics, Howard University, Washington, DC, USA

R. R. Bhargava Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, India

John C. Bowman University of Alberta, Edmonton, AB, Canada

M. S. Bruzon University of Cadiz Puerto Real, Puerto Real, Spain

Wesley S. Burr Queen's University, Kingston, ON, Canada

Giuseppe Campolieti Wilfrid Laurier University, Waterloo, ON, Canada

Cristina Canepa Faculty of Mathematics, University of Bucharest, Bucharest, Romania

Ricardo Castellanos Florida Atlantic University, Boca Raton, FL, USA

P. Chen Waterloo Institute for Nanotechnology, University of Waterloo, Waterloo, ON, Canada

Ludovic Chretien Regal Beloit Corporation, Fort Wayne, IN, USA

David Clendenen Regal Beloit Corporation, Fort Wayne, IN, USA

Rachele Cocks Regal Beloit Corporation, Fort Wayne, IN, USA

Monica G. Cojocaru Mathematics & Statistics, University of Guelph, Guelph, ON, Canada

Morgan Condon Department of Computer Sciences, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV, USA

Ross Cressman Wilfrid Laurier University, Waterloo, ON, Canada

Nareli Cruz-Cortés Centro de Investigación en Computatión, Instituto Politécnico National, Mexico City, Mexico

Attila Dénes Bolyai Institute, University of Szeged, Szeged, Hungary

Eduardo D'Azevedo ORNL, Oak Ridge, TN, USA

T. K. Das Department of Clinical Neurological Sciences, London Health Sciences Centre, Western University, London, ON, Canada

Emre Demircioglu Galatasaray University, Ortakoy, Istanbul, Turkey

Jyotirmoy Deshmukh Toyota Technical Center, Powertrain Control (Model-based Development), Gardena, CA, USA

David H. Dezern Department of Mathematics, Winston-Salem State University, Winston-Salem, NC, USA

R. Dhote University of Toronto and Wilfrid Laurier University, Waterloo, ON, Canada

S.G. Diagne Département de Mathématiques, Université Cheikh Anta Diop, Dakar, Sénégal

J. Dongarra University of Tennessee, Knoxville, TN, USA

Corina S. Drapaca Pennsylvania State University, University Park, PA, USA

G. Duncan Computer Science, Laurentian University, Sudbury, ON, Canada

D. Dutta Physics & Applied Mathematics Unit, Indian Statistical Institute, Kolkata, India

M. R. Ejtehadi Department of Physics, Sharif University of Technology, Tehran, Iran

Nurgun Erdol Florida Atlantic University, Boca Raton, FL, USA

Mujde Erol Genevois Galatasaray University, Ortakoy, Istanbul, Turkey

F. Fillion-Gourdeau Centre de Recherches Mathématiques, Université de Montréal, Montréal, Canada

Igor Leite Freire Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Santo André, SP, Brazil

N. Fugate University of California, Davis, CA, USA

M. Ruiz Galán Department of Applied Mathematics, University of Granada, Granada, Spain

M. L. Gandarias University of Cadiz Puerto Real, Puerto Real, Spain

Mujde Erol Genevois Industrial Engineering Department, Galatasaray University, Istanbul, Turkey

Y. Gningue Department of Mathematics and Computer Sciences, Laurentian University, Sudbury, ON, Canada

H. Gomez University of A Coruña, Coruña, Spain

Álvaro González Department of Mathematics, University of Cadiz, Cadiz, Spain

María Teresa González Montesinos Departamento de Matemática Aplicada I, University of Sevilla, Sevilla, Spain

M. C. Gonzalez Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Ugur Gure Industrial Engineering Department, Galatasaray University, Istanbul, Turkey

Hwashin H. Shin Queen's University, Kingston, ON, Canada

G. B. Hall ESRI, Toronto, ON, Canada

David Hamilton Dalhousie University, Halifax, NS, Canada

C. Hogg Mathematics & Statistics, University of Guelph, Guelph, ON, Canada

Julie Horrocks University of Guelph, Guelph, ON, Canada

Zhiang Hu Chinese University of Hong Kong, Shatin, Hong Kong

Mohammed Ibrahim Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON, Canada

Tomasz Imielinski Department of Computer Science, Rutgers University, New Brunswick, NJ, USA

V. In Space and Naval Warfare Systems Center, San Diego, CA, USA

Kamlesh Jangid Indian Institute of Technology Roorkee, Roorkee, India

M. Jog Department of Clinical Neurological Sciences, London Health Sciences Centre, Western University, London, ON, Canada

Darryl Johnson Department of Computer Sciences, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV, USA

Andrew Kail University of Tennessee, Knoxville, TN, USA

T. Kakiashvili Brain Research, Baycrest, Toronto, ON, Canada

James Kapinski Toyota Technical Center, Powertrain Control (Model-based Development), Gardena, CA, USA

Georgiy P. Karev National Centre for Biotechnology Information, Bethesda, MD, USA

Justin A. Kauffman Pennsylvania State University, University Park, PA, USA

Chaudry Masood Khalique Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, Department of Mathematical Sciences, North-West University, Mmabatho, Republic of South Africa

N. P. Khiabani Department of Chemical Engineering, University of Tehran, Tehran, Iran

Waterloo Institute for Nanotechnology, University of Waterloo, Waterloo, ON, Canada

David J. Klinke Department of Chemical Engineering and Mary Babb Randolph Cancer Center, Department of Microbiology, Immunology and Cell Biology, West Virginia University, Morgantown, WV, USA

W. W. Koczkodaj Computer Science, Laurentian University, Sudbury, ON, Canada

Lukáš Krupička Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Prague 6, Czech Republic

H. Kunze Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada

C. Kuusela Mathematics & Statistics, University of Guelph, Guelph, ON, Canada

R. Lacroix Inria and University Pierre et Marie Curie, Paris, France

Meili Li Department of Applied Mathematics, Donghua University, Shanghai, P. R. China

University of Waterloo, Waterloo, ON, Canada

Xiaozhou Li Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands

X. Z. Liu University of Waterloo, Waterloo, ON, Canada

Xinzhi Liu Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

P. Longhini Space and Naval Warfare Systems Center, San Diego, CA, USA

E. Lorin School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

Weixun Lu Department of Geography, University of Victoria, Victoria, BC, Canada

Sirani M. Perera Daytona State College, Daytona Beach, FL, USA

Roman N. Makarov Wilfrid Laurier University, Waterloo, ON, Canada

R. McKenzie Cardiff University, Cardiff, UK

R. Melnik The MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University, Waterloo, ON, Canada

Ralf Meyer Department of Mathematics and Computer Science, Laurentian University, Sudbury, ON, Canada

Isaiah Elvis Mhlanga International Institute for Symmetry Analysis and Mathematical Modelling, Department of Mathematical Sciences, North-West University, Mmabatho, Republic of South Africa

Abdus Sattar Mia University of Saskatchewan, Saskatoon, SK, Canada

Tshepo Edward Mogorosi Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mmabatho, Republic of South Africa

Simon Morgan Los Alamos National Laboratory, Los Alamos, NM, USA

Dimpho Millicent Mothibi Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mmabatho, Republic of South Africa

Ben Muatjetjeja Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mmabatho, Republic of South Africa

Paul Muir Saint Mary's University, Halifax, NS, Canada

Juan Murcia Instituto Eduardo Torroja, CSIC/IETcc-CSIC, Madrid, Spain

P. Nagarani The University of the West Indies, Kingston, Jamaica W. I.

N. Nedialkov McMaster University, Hamilton, ON, Canada

Nedialko S. Nedialkov Department of Computing and Software, McMaster University, Hamilton, ON, Canada

Nathaniel K. Newlands Science and Technology, Agriculture and Agri-Food Canada, Lethbridge Research Centre, Lethbridge, AB, Canada

Kaya Ocakoglu Industrial Engineering Department, Galatasaray University, Istanbul, Turkey

Vadim Olshevsky University of Connecticut, Storrs, CT, USA

Francisco Ortegón Gallego Departamento de Matemáticas, University of Cádiz, Cádiz, Spain

A. Palacios Nonlinear Dynamical Systems Group, Department of Mathematics, San Diego State University, San Diego, CA, USA

Sudhakar G. Pandit Department of Mathematics, Winston-Salem State University, Winston-Salem, NC, USA

Lev F. Petrov Russian Plekhanov University of Economics, Moscow, Russia

National Research University Higher School of Economics, Moscow, Russia

Jack Pew Saint Mary's University, Halifax, NS, Canada

Dan Pirjol J. P. Morgan, New York, NY, USA

Traian A. Pirvu Math and Stat Department, McMaster University, Hamilton, ON, Canada

Tracy A. Porcelli Physicist/Scientific Consultant, 316 7th Avenue South, Lethbridge, AB, Canada

P. Pourafshary Petroleum and chemical engineering department, Sultan Qaboos University, Muscat, Oman

J. Pryce Cardiff University, Cardiff, UK

John D. Pryce Cardiff School of Mathematics, Cardiff University, Cardiff, UK

Gergely Röst Bolyai Institute, University of Szeged, Szeged, Hungary

N. Rajakumar Department of Anatomy and Cell Biology, Western University, London, ON, Canada

Luis A. Rivera-Zamarripa Centro de Investigación en Computación del Instituto Politécnico Nacional, México City, México

Malcolm Roberts University of Strasbourg, Strasbourg, France

Steven A. Roberts Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, ON, Canada

M. Rosa University of Cadiz Puerto Real, Puerto Real, Spain

Matthew Rueffer University of Guelph, Guelph, ON, Canada

Jennifer K. Ryan School of Mathematics, University of East Anglia, Norwich, UK

Carolyn Salafia Placental Analytics, LLC, Larchmont, NY, USA

Institute for Basic Research, Staten Island, NY, USA

Armaghan Salehian Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON, Canada

K. Sarikhani Waterloo Institute for Nanotechnology, University of Waterloo, Waterloo, ON, Canada

B. T. Sebastian The University of the West Indies, Kingston, Jamaica W. I.

University of Technology, Kingston, Jamaica

Schehrazad Selmane LIFORCE, Faculty of Mathematics, The University of Science and Technology Houari Boumediene, Algiers, Algeria

R. Shaffer Nonlinear Dynamical Systems Group, Department of Mathematics, San Diego State University, San Diego, CA, USA

Hwashin H. Shin Queen's University, Kingston, ON, Canada

Hasan Shodiev Wilfrid Laurier University, Waterloo, ON, Canada

Alla Shymanska School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

Priscila Leal da Silva Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Santo André, SP, Brazil

Andrew Skelton Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada

Arash Soleimani Dahaj University of Waterloo, Waterloo, ON, Canada

M. Soltani Waterloo Institute for Nanotechnology, University of Waterloo, Waterloo, ON, Canada

Department of Mechanical Engineering, K. N. T. University of Technology, Tehran, Iran

Division of Nuclear Medicine, Johns Hopkins University, School of Medicine, Baltimore, MD, USA

Peter Stechlinski University of Waterloo, Waterloo, ON, Canada

Shiquan Su University of Tennessee, Knoxville, TN, USA

Winston L. Sweatman Centre for Mathematics in Industry, Institute of Natural and Mathematical Sciences, Massey University, Albany, Auckland, New Zealand

P. M. Takouda School of Commerce & Administration, Laurentian University, Sudbury, Canada

Guangning Tan Department of Computing and Software, McMaster University, Hamilton, Canada

Fairouz Tchier Mathematics Department, Riyadh, Saudi Arabia

E. W. Thommes Mathematics & Statistics, University of Guelph, Guelph, ON, Canada

Edward Thommes GlaxoSmithKline, Mississauga, ON, Canada

Jungang Tian Department of Applied Mathematics, Donghua University, Shanghai, P. R. China

D. La Torre Department of Applied Mathematics and Sciences, Khalifa University, Abu Dhabi, UAE

Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy

Tan Tran Wilfrid Laurier University, Waterloo, ON, Canada

J. Turtle Nonlinear Dynamical Systems Group, Department of Mathematics, San Diego State University, San Diego, CA, USA

Ziya Ulukan Galatasaray University, Ortakoy, Istanbul, Turkey

Pooja Raj Verma Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, India

Giuseppe Viglialoro Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

Mark P. Wachowiak Department of Computer Science and Mathematics, Nipissing University, North Bay, ON, Canada

Renata Wachowiak-Smolíková Department of Computer Science and Mathematics, Nipissing University, North Bay, ON, Canada

Jie Yu Wang Department of Mathematics and Computer Science, Laurentian University, Sudbury, ON, Canada

Qing Wang Department of Computer Sciences, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV, USA

Zhijun Wang Department of Computer Sciences, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV, USA

Joshua Westhoff Regal Beloit Corporation, Fort Wayne, IN, USA

Marianne Wilcox University of Guelph, Guelph, ON, Canada

Allan R. Willms Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada

Kwai Wong University of Tennessee, Knoxville, TN, USA

Sarah Wong Dalhousie University, Halifax, NS, Canada

Qinglan Xia Department of Mathematics, University of California at Davis, Davis, CA, USA

H. Xiao Department of Computer Science, University of California, Davis, CA, USA

W.-C. Xie University of Waterloo, Waterloo, ON, Canada

Wei-Chau Xie Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada

John Robert Yaros Department of Computer Science, Rutgers University, New Brunswick, NJ, USA

Peter J. S. Young NCI Agency, The Hague, AK, The Netherlands

Hongtao Zhang University of Waterloo, Waterloo, ON, Canada

Kexue Zhang Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

Hanqi Zhuang Florida Atlantic University, Boca Raton, FL, USA

Jonathan Zimmerling Department of Computer Science and Mathematics, Nipissing University, North Bay, ON, Canada

Stroma Service Consulting Ltd., North Bay, ON, Canada

J. Zu University of Toronto, Toronto, ON, Canada

Exact Solutions and Conservation Laws of the Joseph-Egri Equation with Power Law Nonlinearity

Abdullahi Rashid Adem and Chaudry Masood Khalique

Abstract In this chapter we obtain exact solutions of the Joseph-Egri equation with power law nonlinearity, which arises in various problems in many scientific applications. The Lie group analysis and simplest equation method are used to carry out the integration of this equation. The solutions obtained are travelling wave solutions. Moreover, the conservation laws for the Joseph-Egri equation with power law nonlinearity are constructed by using the multiplier method.

1 Introduction

Nonlinear differential equations and, in particular, nonlinear evolution equations (NLEEs) are widely used as models to describe physical phenomena in many fields of science. Therefore, it is imperative that their solutions be found. However, finding exact solutions of NLEEs is not an easy task and during the past few decades, many methods have been developed by researchers to find explicit solutions for such equations. Some of the methods commonly used in the literature are the inverse scattering transform method, the Lie group method, the variational iteration method, the exp-function method, the sine-cosine method and the (G'/G)-expansion method. See, for example, [1] and references therein.

In this chapter we study one such NLEE, namely the Joseph-Egri equation with power law nonlinearity that is given by

$$u_t + u_x + \alpha u^n u_x + \beta u_{xtt} = 0. \tag{1}$$

C. M. Khalique (✉)
Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, Department of Mathematical Sciences, North-West University, Mmabatho, Republic of South Africa
e-mail: Masood.Khalique@nwu.ac.za

A. R. Adem
International Institute for Symmetry Analysis and Mathematical Modelling,
North-West University, Mafikeng Campus, Private Bag X 2046,
2735, Mmabatho, South Africa
e-mail: Abdullahi.R.Adem@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_1

Here, in (1) α, β and n are real valued nonzero constants. For $n = 1$, a 1-soliton and 2-soliton solutions have been found in [8] for the Joseph-Egri equation. See also [4].

In this chapter, Lie group analysis [5, 7] in conjunction with the simplest equation method [6] is employed to obtain some exact solutions of (1). In addition to this, conservation laws will be derived for (1) using the multiplier method [2].

2 Exact Solutions Using Lie Group Analysis

First we find the Lie point symmetries of (1) and latter use them to construct exact solutions.

2.1 Lie Point Symmetries

A Lie point symmetry of a partial differential equation (PDE) is an invertible transformation of the dependent and independent variables that leaves the equation unchanged. The symmetry group of the Joseph-Egri equation with power law nonlinearity (1) will be generated by the vector field of the form

$$\Gamma = \xi^1(t,x,u)\frac{\partial}{\partial t} + \xi^2(t,x,u)\frac{\partial}{\partial x} + \eta(t,x,u)\frac{\partial}{\partial u}. \tag{2}$$

Applying the third prolongation $\mathrm{pr}^{(3)}\Gamma$ [7] to (1), we obtain an overdetermined system of linear PDEs. Then solving this resultant system of linear PDEs one obtains the following two Lie point symmetries:

$$\Gamma_1 = \frac{\partial}{\partial t}, \quad \Gamma_2 = \frac{\partial}{\partial x}.$$

2.2 Exact Solutions

We now use the two symmetries to obtain exact solution of (1). The combination $\Gamma_1 + \nu\Gamma_2$, of the two symmetries Γ_1 and Γ_2 yields the two invariants

$$z = x - \nu t, \quad F = u,$$

which gives rise to a group invariant solution $F = F(z)$ and then using these invariants, (1) is transformed into the nonlinear third-order ordinary differential equation (ODE)

$$\beta\nu^2 F'''(z) + \alpha F(z)^n F'(z) - \nu F'(z) + F'(z) = 0. \tag{3}$$

Integrating the above equation and taking the constant of integration to be zero we obtain the second-order ODE

$$\beta \nu^2 F''(z) + \frac{\alpha F(z)^{n+1}}{n+1} + (1-\nu)F(z) = 0. \tag{4}$$

Now multiplying Eq. (4) by F', integrating once and taking the constant of integration to be zero, we obtain the first-order variable separable ODE

$$\frac{1}{2}\left\{\beta \nu^2 F'(z)^2 - (\nu-1)F(z)^2\right\} + \frac{\alpha F(z)^{n+2}}{(n+1)(n+2)} = 0. \tag{5}$$

Integrating and reverting back to the original variables, we obtain the following group-invariant solution of (1), for arbitrary values of n, in the form:

$$u(x,t) = \left[P_1\left\{\tanh^2\left(\frac{1}{2}(\pm P_2(x-\nu t)+P_3)\right) - 1\right\}\right]^{1/n}, \tag{6}$$

where

$$P_1 = -\frac{(1+n)(2+n)(-1+\nu)}{2\alpha},$$
$$P_2 = \frac{n\sqrt{\beta(\nu-1)}}{\beta\nu},$$
$$P_3 = -n\sqrt{(1+n)(2+n)(-1+\nu)}C.$$

3 Exact Solutions Using Simplest Equation Method

In this section we use the simplest equation method, which was introduced by Kudryashov [6] to solve the third-order ODE (3). We will use the Bernoulli and Ricatti equations as our simplest equations. It is well-known that their solutions can be written in elementary functions [1].

Let us consider the solution of (3) in the form

$$F(z) = \sum_{i=0}^{M} A_i (H(z))^i, \tag{7}$$

where $H(z)$ satisfies the Bernoulli or the Ricatti equation, M is a positive integer that can be determined by balancing procedure and $A_0, \cdots, A_M$ are parameters to be determined.

3.1 Solutions of (1) Using the Bernoulli Equation as the Simplest Equation

The balancing procedure, in this case, yields $M = 2$ so the solutions of (3) are of the form

$$F(z) = A_0 + A_1 H + A_2 H^2. \tag{8}$$

Substituting (8) into (3) and making use of the Bernoulli equation

$$H'(z) = aH(z) + bH^2(z)$$

and then equating all coefficients of the functions H^i to zero, we obtain an algebraic system of equations in terms of A_0, A_1 and A_2. On solving this system of algebraic equations, with the help of Mathematica, we obtain

$$\begin{aligned} A_0 &= \frac{-a^2\beta\nu^2 + \nu - 1}{\alpha},\\ A_1 &= -\frac{12ab\beta\nu^2}{\alpha},\\ A_2 &= -\frac{12b^2\beta\nu^2}{\alpha}. \end{aligned}$$

As a result, a solution of (1) is

$$\begin{aligned} u(t,x) = A_0 + A_1 a &\left\{\frac{\cosh[a(z+C)] + \sinh[a(z+C)]}{1 - b\cosh[a(z+C)] - b\sinh[a(z+C)]}\right\}\\ &+ A_2 a^2 \left\{\frac{\cosh[a(z+C)] + \sinh[a(z+C)]}{1 - b\cosh[a(z+C)] - b\sinh[a(z+C)]}\right\}^2, \end{aligned} \tag{9}$$

where $z = x - \nu t$ and C is a constant of integration.

3.2 Solutions of (1) Using Riccati Equation as the Simplest Equation

The balancing procedure yields $M = 2$ so the solutions of (3) are of the form

$$F(z) = A_0 + A_1 H + A_2 H^2. \tag{10}$$

Substituting (10) into (3) and making use of the Riccati equation

$$H'(z) = aH^2(z) + bH(z) + c,$$

we obtain a system of algebraic equations in terms of A_0, A_1, A_2, as before. Solving these algebraic equations, with the aid of Mathematica, we obtain

$$\begin{aligned} A_0 &= \frac{-8a\beta c\nu^2 - b^2\beta\nu^2 + \nu - 1}{\alpha}, \\ A_1 &= -\frac{12ab\beta\nu^2}{\alpha}, \\ A_2 &= -\frac{12a^2\beta\nu^2}{\alpha}. \end{aligned}$$

Thus, as a result, solutions of (1) are

$$\begin{aligned} u(t,x) = A_0 + A_1&\Big\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\Big[\frac{1}{2}\theta(z+C)\Big]\Big\} \\ &+ A_2\Big\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\Big[\frac{1}{2}\theta(z+C)\Big]\Big\}^2 \end{aligned} \tag{11}$$

and

$$\begin{aligned} u(t,x) = A_0 + A_1&\Big\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\Big(\frac{1}{2}\theta z\Big) + \frac{\operatorname{sech}\big(\frac{\theta z}{2}\big)}{C\cosh\big(\frac{\theta z}{2}\big) - \frac{2a}{\theta}\sinh\big(\frac{\theta z}{2}\big)}\Big\} \\ &+ A_2\Big\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\Big(\frac{1}{2}\theta z\Big) + \frac{\operatorname{sech}\big(\frac{\theta z}{2}\big)}{C\cosh\big(\frac{\theta z}{2}\big) - \frac{2a}{\theta}\sinh\big(\frac{\theta z}{2}\big)}\Big\}^2, \end{aligned} \tag{12}$$

where $z = x - \nu t$ and C is a constant of integration.

3.3 Construction of Conservation Laws for (1)

We now construct conservation laws for the Joseph-Egri equation with power law nonlinearity (1) in this section. The multiplier method will be employed [2]. See also [3]. The zeroth-order multiplier for (1) is, $\Lambda(t,x,u)$ that is given by

$$\Lambda = C_1 u + C_2$$

where C_1 and C_2 are arbitrary constants. Thus, corresponding to the above multiplier we have the following two conserved vectors:

$$\begin{aligned} \Phi_1^t &= \frac{1}{6}\Big\{4\beta u u_{tx} + 3u^2 - 2\beta u_t u_x\Big\}, \\ \Phi_1^x &= \frac{1}{6(n+2)}\Big\{2\beta n u_{texttt}u + 4\beta u_{texttt}u + 6\alpha u^{n+2} + 3nu^2 + 6u^2 - \beta n u_t^2 - 2\beta u_t^2\Big\} \end{aligned}$$

and

$$\Phi_2^t = \frac{1}{3}\Big\{3u + 2\beta u_{tx}\Big\},$$
$$\Phi_2^x = \frac{1}{3(n+1)}\Big\{3\alpha u^{n+1} + 3nu + 3u + \beta n u_{texttt} + \beta u_{texttt}\Big\}.$$

4 Concluding Remarks

In this chapter we obtained the exact solutions of the Joseph-Egri equation with power law nonlinearity by employing the Lie group analysis and the simplest equation method. Moreover, we also derived the conservation laws for the underlying equation by using the multiplier method.

Acknowledgement Abdullahi Rashid Adem thanks the NRF for financial support. Chaudry Masood Khalique would like to thank the Organizing Committee of "International Conference: AMMCS-2013," Waterloo, Canada for their kind hospitality during the conference.

References

1. Adem, A.R., Khalique, C.M.: Symmetry reductions, exact solutions and conservationlaws of a new coupled KdV system. Commun. Nonlinear Sci. Numer. Simul. **17,** 3465–3475 (2012)
2. Anco, S.C., Bluman, G.W.: Direct construction method for conservation laws of partial differential equations. Part I: examples of conservation law classifications. Eur. J. Appl. Math. **13,** 545–566 (2002)
3. Anthonyrajah, M., Mason, D.P.: Conservation laws and invariant solutions in the Fanno model for turbulent compressible flow. Math. Comput. Appl. **15,** 529–542 (2010)
4. Hereman, W., Banerjee, P.P., Korpel, A., Assanto, G., van Immerzeele, A., Meerpole, A.: Exact solitary wave solutions of nonlinear evolution and wave equations using a direct algebraic method. J. Phys. A. Math. Gen. **1,** 607–628 (1986)
5. Ibragimov, N.H.: CRC Handbook of Lie Group Analysis of Differential Equations, vols. 1–3, CRC Press, Boca Raton (1994–1996)
6. Kudryashov, N.A.: Simplest equation method to look for exact solutions of nonlinear differential equations. Chaos Solitons Fractals **24,** 1217–1231 (2005)
7. Olver, P.J.: Applications of Lie Groups to Differential Equations, Graduate Texts in Mathematics, 107, 2nd edn. Springer-Verlag, Berlin (1993)
8. Taghizadeh, N., Mirzazadeh, M.: The multisoliton solutions of some nonlinear partial differential equations. Appl. Appl. Math. **6,** 284–291 (2011)

ML-α-Deconvolution Model in a Bounded Domain with a Vertical Regularization

Hani Ali

Abstract In this chapter, we consider the deconvolution modified Leray alpha (ML-α-deconvolution) model with fractional filter acting only in one variable $\mathbb{A}_{3,\theta} = I + \alpha_3^{2\theta}(-\partial_3)^{2\theta}$, where $0 \leq \theta \leq 1$ controls the degree of smoothing in the filter. We study the global existence and uniqueness of solutions to the vertical ML-α-deconvolution model on a bounded product domain of the type $D = \Omega \times (-\pi, \pi)$, where Ω is a smooth domain with homogeneous Dirichlet boundary conditions on the boundary $\partial\Omega \times (-\pi, \pi)$, and with periodic boundary conditions in the vertical variable. To present the model, we define the vertical Nth Van Cittert deconvolution operator by $D_{N,\theta} = \sum_{i=0}^{N} (I - \mathbb{A}_{3,\theta}^{-1})^i$. The vertical ML-$\alpha$-deconvolution model is then defined by replacing the nonlinear term in the Navier–Stokes equations $(v \cdot \nabla)v$ by $(v \cdot \nabla)D_{N,\theta}(\overline{v})$ where v is the velocity, and $\overline{v} = \mathbb{A}_{3,\theta}^{-1}(v)$ is the smoothed velocity. We adapt the ideas from (H. Ali, Approximate Deconvolution Model in a bounded domain with a vertical regularization. J Math Anal Appl **408**, 355–363 (2013)) to prove that the vertical ML-α-deconvolution model which is derived by using $\mathbb{A}_{3,\theta}$, has a unique weak solution for any $\theta > \frac{1}{2}$.

1 Introduction

In this chapter, we consider the deconvolution modified Leray alpha (ML-α-deconvolution) model with fractional filter acting only in one variable

$$\mathbb{A}_{3,\theta} := I + \alpha_3^{2\theta}(-\partial_3)^{2\theta}, \quad 0 \leq \theta \leq 1, \tag{1}$$

where θ controls the degree of smoothing in the filter.

This filter is less memory consuming than the classical one (see, e.g., [3, 5, 7, 8]). Moreover, there is no need to introduce artificial boundary conditions for Helmholtz operator. It was shown in [4] that the Large Eddy Simulation models which are derived by using $\mathbb{A}_{3,\theta}$ for any $\theta > \frac{1}{2}$, are well posed. Motivated by this work [4],

H. Ali (✉)
Institut Jean le Rond d'Alembert, CNRS & UPMC Université Paris 06,
UMR 7190, 75005 Paris, France
e-mail: hani.ali@etu.upmc.fr

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_2

we study the global existence and uniqueness of solutions to the vertical ML-α-deconvolution model on a bounded product domain of the type $D = \Omega \times (-\pi, \pi)$, where Ω is a smooth domain with homogeneous Dirichlet boundary conditions on the boundary $\partial\Omega \times (-\pi, \pi)$, and with periodic boundary conditions in the vertical variable. To present the model, we define the vertical Nth Van Cittert deconvolution operator by

$$D_{N,\theta} = \sum_{i=0}^{N} (I - \mathbb{A}_{3,\theta}^{-1})^i. \tag{2}$$

The vertical ML-α-deconvolution model is then defined, for some fixed $\theta > 0$, with a filtering radius $\alpha_3 > 0$, a kinematic viscosity $\nu > 0$, a deconvolution order $N \geq 0$, and an initial velocity v_0 as follows,

$$\partial_t v + (v \cdot \nabla) D_{N,\theta}(\overline{v}) - \nu \Delta v + \nabla p = f, \tag{3}$$

$$\nabla \cdot v = 0, \tag{4}$$

$$v(0) = v_0, \tag{5}$$

where v and p are the velocity and the pressure, $\overline{v} = \mathbb{A}_{3,\theta}^{-1}(v)$ is the smoothed velocity, and f is a forcing term.

For simplicity, we consider the domain $D = \{x \in \mathbb{R}^3, x_1^2 + x_2^2 < d, \ -\pi < x_3 < \pi\}$ with 2π periodicity with respect to x_3. Therefore, the deconvolution model in this chapter is chosen to model the flow through a cylinder or a pipe with periodic boundary conditions with respect to x_3. We note that the filter is acting only in the vertical variable, that is why it is possible to require the periodicity only in x_3. Moreover, we consider the unfiltered function with homogeneous Dirichlet boundary conditions on the boundary $\partial D = \partial\Omega \times (-\pi, \pi)$. These boundary conditions of the unfiltered function are supposed to be the same as the filtered ones, in order to prevent from introducing artificial boundary conditions. In order to state our main result, let us define the following spaces:

$$L^2(D) := \left\{v \in L^2(D)^3, 2\pi\text{-periodic in } x_3\right\}, \tag{6}$$

$$H := \left\{v \in L^2(D), \text{ such that } \nabla \cdot v = 0 \text{ and } v \cdot n = 0 \text{ on } \partial\Omega \times (-\pi, \pi)\right\}, \tag{7}$$

$$V := \left\{v \in H, \text{ such that } \nabla v \in L^2(D) \text{ and } v = 0 \text{ on } \partial\Omega \times (-\pi, \pi)\right\}. \tag{8}$$

Next, we give a definition of what is called a weak solution to the vertical ML-α-deconvolution model.

Definition 1 Let $f \in L^2(0, T; H)$ and $v_0 \in H$. For any $0 \leq \theta \leq 1$ and $0 \leq N < \infty$, the couple (v, p) is called a weak solution to (3)–(5) if

$$v \in \mathcal{C}_w(0, T; H) \cap L^2(0, T; V), \tag{9}$$

and the couple (v, p) fulfills

$$\begin{aligned}&\int_0^T \langle \partial_t v, \varphi\rangle - \langle D_{N,\theta}(\overline{v}) \otimes v, \nabla\varphi\rangle + \nu\langle \nabla v, \nabla\varphi\rangle + \langle \nabla p, \varphi\rangle \, dt \\ &\quad = \int_0^T \langle f, \varphi\rangle \, dt \qquad \text{for all } \varphi \in \mathcal{C}_c^\infty([0,T] \times D).\end{aligned} \tag{10}$$

Moreover,

$$v(0) = v_0. \tag{11}$$

Our main result is the following.

Theorem 1 *Assume $f \in L^2(0,T;H)$ and $v_0 \in H$. let $0 \le N < \infty$ be a given and fixed number and let $\theta > \frac{1}{2}$. Then problem (3)–(5) has a unique weak solution.*

This result holds also true on the whole space $\mathbb{R}^3$ and on the torus $\mathbb{T}_3$. The vertical ML-α-deconvolution with $N = 0$ becomes the modified Leray alpha (ML-α) model of turbulence [2, 6] with filter acting only in one variable. Consequently, Theorem 1 gives us also existence and uniqueness of solutions to the vertical ML-α model of turbulence on the bounded domain D. Other α models, with partial filter, will be reported in a forthcoming paper.

2 Notation and Auxiliary Result

In this section, we introduce relevant function spaces and we recall an auxiliary result used in the proof of the main result.

Let $1 < p \le +\infty$ and $1 < q \le +\infty$. We denote by $L_v^q L_h^p(D) = L^q((-\pi,+\pi); L^p(\Omega))$ the space of functions g such that $(\int_{-\pi}^{+\pi}(\int_\Omega |g(x_1,x_2,x_3)|^p dx_1 dx_2)^{q/p} dx_3)^{1/q} < +\infty$.

We denote by $\|v\|_2 := \int_D v \cdot v \, dx$ the usual norm in $L^2(D)^3$.

The following lemma will play an important role [4].

Lemma 1 *There exists a positive constant $C > 0$ such that, for any $s > \frac{1}{2}$ and for any smooth enough divergence-free vector fields u, v, and w, the following estimate holds,*

$$|((u \cdot \nabla)v, w)| \le C\|u\|_2^{\frac{1}{2}} \|\nabla u\|_2^{\frac{1}{2}} \left(\|\nabla v\|_2^{1-\frac{1}{2s}} \|\partial_3^s \nabla v\|_2^{\frac{1}{2s}} + \|\nabla v\|_2 \right) \|w\|_2^{\frac{1}{2}} \|\nabla w\|_2^{\frac{1}{2}}.$$

3 The Vertical Filter and the Vertical Deconvolution Operator

In this section, we record some properties of the vertical filter and of the vertical deconvolution operator. Let v be a smooth function of the form $v = \sum_{k_3 \in \mathbb{Z}\setminus\{0\}} c_{k_3}(x_1,x_2) e^{\mathrm{i} k_3 x_3}$. The action of the vertical filter on $v(x) =$

$\sum_{k_3\in\mathbb{Z}\setminus\{0\}} c_{k_3}(x_1,x_2)e^{\mathrm{i}\,k_3x_3}$ can be written as $\mathbb{A}_{3,\theta}(v) = \sum_{k_3\in\mathbb{Z}\setminus\{0\}} \mathcal{A}_\theta(k_3)c_{k_3}(x_1,x_2)$ $e^{\mathrm{i}\,k_3x_3}$, where the symbol with respect to x_3 of the vertical filter is given by

$$\mathcal{A}_\theta(k_3) = \left(1+\alpha^{2\theta}|k_3|^{2\theta}\right). \tag{12}$$

Therefore, by using the Parseval's identity with respect to x_3 we get,

$$\|\mathbb{A}_{3,\theta}^{\frac{1}{2}}v\|_2^2 = \|v\|_2^2 + \alpha^{2\theta}\|\partial_3^\theta v\|_2^2 = \left(\mathbb{A}_{3,\theta}v, v\right). \tag{13}$$

The deconvolution operator $D_{N,\theta} = \sum_{i=0}^N (I-\mathbb{A}_{3,\theta}^{-1})^i$ is constructed by using the vertical filter with fractional regularization (1). For a fixed $N > 0$ and for $\theta = 1$, we recover a vertical operator form from the Van Cittert deconvolution operator.

A straightforward calculation yields

$$D_{N,\theta}\left(\sum_{k_3\in\mathbb{Z}\setminus\{0\}} c_{k_3}(x_1,x_2)e^{\mathrm{i}\,k_3x_3}\right) = \sum_{k_3\in\mathbb{Z}\setminus\{0\}} \mathcal{D}_{N,\theta}(k_3)c_{k_3}(x_1,x_2)e^{\mathrm{i}\,k_3x_3}, \tag{14}$$

where for $k_3 \in \mathbb{Z}\setminus\{0\}$ and $\theta \geq 0$, $\mathcal{D}_{N,\theta}(k_3)$ verifies:

$$\mathcal{D}_{0,\theta}(k_3) = 1, \tag{15}$$

$$1 \leq \mathcal{D}_{N,\theta}(k_3) \leq N+1 \quad \text{for each } N > 0, \tag{16}$$

$$\text{and } \mathcal{D}_{N,\theta}(k_3) \leq \mathcal{A}_{3,\theta} \quad \text{for a fixed } \alpha > 0. \tag{17}$$

From the previous hypothesis, one can prove the following Lemma by adapting the results summarized in the isotropic case in [1]:

Lemma 2 *For all $s \geq -1$, $\theta \geq 0$, $k_3 \in \mathbb{Z}\setminus\{0\}$ and for each $N > 0$, there exists a constant $C > 0$ such that for all v sufficiently smooth we have*

$$\|v\|_{s,2} \leq \|D_{N,\theta}\,(v)\,\|_{s,2} \leq (N+1)\|v\|_{s,2}, \tag{18}$$

$$\|\mathbb{A}_{3,\theta}{}^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(v)\|_{s,2} \leq \|v\|_{s,2}, \tag{19}$$

$$\|\mathbb{A}_{3,\theta}^{-\frac{1}{2}}(v)\|_{s,2}^2 = \|\overline{v}\|_{s,2}^2 + \alpha_3^{2\theta}\|\partial_3^\theta\overline{v}\|_{s,2}^2. \tag{20}$$

4 Sketch of the Proof of the Main Result

We briefly present the main ideas of the proof of Theorem 1. The proof follows from the following a priori estimates with a Galerkin method.

For further information, we refer the reader to [1, 4] and the references therein.

Proof Multiplying (3) with $D_{N,\theta}(\overline{v})$ integrating over time from 0 to t, for all $t \in [0,T]$, and using standard manipulations lead to the a priori estimate

$$\begin{aligned}\frac{1}{2}\|\mathbb{A}_\theta^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(v)\|_2^2 + \nu\int_0^t \|\nabla\mathbb{A}_\theta^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(v)\|_2^2\,ds \\ = \int_0^t \langle\mathbb{A}_\theta^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(f), \mathbb{A}_\theta^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(v)\rangle\,ds + \frac{1}{2}\|\mathbb{A}_\theta^{-\frac{1}{2}}D_{N,\theta}^{\frac{1}{2}}(v_0)\|_2^2.\end{aligned} \tag{21}$$

By using the duality norm combined with Young inequality and inequality (19), we conclude from (21) that

$$\begin{aligned}\sup_{t\in[0,T]} \|\mathbb{A}_\theta^{-\frac{1}{2}} D_{N,\theta}^{\frac{1}{2}}(v)\|_2^2 + \nu \int_0^T \|\nabla \mathbb{A}_\theta^{-\frac{1}{2}} D_{N,\theta}^{\frac{1}{2}}(v)\|_2^2 \, dt \\ \le \|v_0\|_2^2 + \frac{C}{\nu} \int_0^T \|f\|_2^2 \, dt.\end{aligned} \tag{22}$$

We deduce from (22) and (20) that

$$\overline{v} \text{ and } \partial_3^\theta \overline{v} \in L^\infty(0,T;H) \cap L^2(0,T;V). \tag{23}$$

Thus, it follows from (18) that

$$D_{N,\theta}(\overline{v}^n) \text{ and } \partial_3^\theta D_{N,\theta}(\overline{v}^n) \in L^\infty(0,T;H) \cap L^2(0,T;V). \tag{24}$$

Multiplying (3) with v we conclude that

$$\frac{1}{2}\frac{d}{dt}\|v\|_2^2 + \nu\|\nabla v\|_2^2 \le \left|\left((v\cdot\nabla)D_{N,\theta}(\overline{v}), v\right)\right| + |\langle f, v\rangle| \,. \tag{25}$$

For $\theta > \frac{1}{2}$ we have

$$\begin{aligned}\left|\left((v\cdot\nabla)D_{N,\theta}(\overline{v}), v\right)\right| &\le C\|v\|_2\|\nabla v\|_2 \\ &\quad \times \left(\|\nabla D_{N,\theta}(\overline{v})\|_2^{1-\frac{1}{2\theta}} \|\partial_3^\theta \nabla D_{N,\theta}(\overline{v})\|_2^{\frac{1}{2\theta}} + \|\nabla D_{N,\theta}(\overline{v})\|_2 \right) \\ &\le C \left(\|\nabla D_{N,\theta}(\overline{v})\|_2^{2-\frac{1}{\theta}} \|\partial_3^\theta \nabla D_{N,\theta}(\overline{v})\|_2^{\frac{1}{\theta}} + \|\nabla D_{N,\theta}(\overline{v})\|_2^2 \right) \\ &\quad \times \|v\|_2^2 + \frac{\nu}{4}\|\nabla v\|_2^2,\end{aligned} \tag{26}$$

where we have used Lemma 1 and the Young inequality.

The second term in right hand side of (25) is estimated by

$$|\langle f, v\rangle| \le C\|f\|_2\|\nabla v\|_2 \le C\|f\|_2^2 + \frac{\nu}{4}\|\nabla v\|_2^2. \tag{27}$$

Thus, (26) and (27) lead to the conclusion that

$$\begin{aligned}&\frac{d}{dt}\|v\|_2^2 + \nu\|\nabla v\|_2^2 \le \\ &\quad C\left(\|\nabla D_{N,\theta}(\overline{v})\|_2^{2-\frac{1}{\theta}} \|\partial_3^\theta \nabla D_{N,\theta}(\overline{v})\|_2^{\frac{1}{\theta}} + \|\nabla D_{N,\theta}(\overline{v})\|_2^2 \right) \|v\|_2^2 + C\|f\|_2^2.\end{aligned} \tag{28}$$

Integrating (28) over time from 0 to T and using Gronwall's Lemma and (24) lead to the following estimate

$$\sup_{t\in[0,T]} \|v\|_2^2 + \nu \int_0^T \|\nabla v\|_2^2 \, dt \le C. \tag{29}$$

We deduce from (29) that

$$v \in L^{\infty}(0,T;H) \cap L^{2}(0,T;V). \tag{30}$$

Finally, we check the question of the uniqueness of the solution. Let $\theta > \frac{1}{2}$ and let (v_1, p_1) and (v_2, p_2) be any weak solutions of (3)–(5) on the interval $[0, T]$, with initial values $v_1(0)$ and $v_2(0)$. Let us denote by $\delta v = v_2 - v_1$ and $\delta D_{N,\theta}(\overline{v}) = D_{N,\theta}(\overline{v}_2) - D_{N,\theta}(\overline{v}_1)$. We subtract the equation for v_1 from the equation for v_2 and test it with δv, we formally get:

$$\begin{aligned}&\frac{d}{dt}\|\delta v\|_2^2 + \nu\|\nabla\delta v\|_2^2 \\ &\quad \le C\|\nabla D_{N,\theta}(\overline{v}_1)\|_2^{2-\frac{1}{\theta}}\|\partial_3^{\theta}\nabla D_{N,\theta}(\overline{v}_1)\|_2^{\frac{1}{\theta}}\|\delta v\|_2^2 + C\|v_2\|_2^2\|\nabla v_2\|_2^2\|\delta v\|_2^2\end{aligned} \tag{31}$$

where we have used Lemma 1, the Young inequality and the fact that $\|\nabla\delta D_{N,\theta}(\overline{v})\|_2 \le C\|\nabla\delta v\|_2$ and $\|\nabla\partial_3^{\theta}\delta D_{N,\theta}(\overline{v})\|_2 \le C\|\nabla\delta v\|_2$.

Since $\|\nabla D_{N,\theta}(\overline{v}_1)\|_2^{2-\frac{1}{\theta}}\|\partial_3^{\theta}\nabla D_{N,\theta}(\overline{v}_1)\|_2^{\frac{1}{\theta}} + \|v_2\|_2^2\|\nabla v_2\|_2^2 \in L^1([0,T])$, we conclude by using Gronwall's inequality the continuous dependence of the solutions on the initial data in the $L^{\infty}(0,T;H)$ norm. In particular, if $\delta v_0 = 0$ then $\delta v = 0$ and the solutions are unique for all $t \in [0, T]$. □

References

1. Ali, H.: On the High Accuracy ML-α-Deconvolution Turbulence Model. Under revision in Analysis and Applications
2. Ali, H.: Ladder theorem and length scale estimates for the modified Leray-alpha model of turbulence. Commun. Math. Sci. **2,** 477–491 (2012)
3. Ali, H.: On a critical Leray-α model of turbulence. Nonlinear Anal.: Real World Appl. **14,** 1536–1584 (2013)
4. Ali, H.: Approximate deconvolution model in a bounded domain with vertical regularization. J. Math. Anal. Appl. **408,** 355–363 (2013)
5. Ali, H.: Theory for the rotational deconvolution model of turbulence with fractional regularization. Appl. Anal. **93,** 339–355 (2014)
6. Ilyin, A.A., Lunasin, E.M., Titi, E.S.: A modified Leray-alpha subgrid-scale model of turbulence. Nonlinearity **19,** 879–897 (2006)
7. Layton, W., Lewandowski, R.: A high accuracy Leray-deconvolution model of turbulence and its limiting behavior. Anal. Appl. **6,** 1–27 (2008)
8. Neda, M., Rebholz, L.G., Layton, W., Manica, C.C.: Numerical analysis and computational testing of a high-accuracy Leray-deconvolution model of turbulence. Numer. Methods Part. Differ. Equ. **24,** 555–582 (2008)

Solving the Linear Transportation Problem by Modified Vogel Method

D. Almaatani, S.G. Diagne, Y. Gningue and P. M. Takouda

Abstract In this chapter, we propose a modification of the Vogel Approximation Method (VAM) used to obtain near optimal solutions to linear transportation problems. This method, called Modified Vogel Method (MVM), consists of performing the row and column reduction of the cost matrix and then applying the classical Vogel method to the equivalent transportation problem with the reduced cost matrix. We prove that when no further reduction of a cost matrix is required, we do obtain an optimal solution, not an approximate one. We identify some cases when such a behavior occurs and provides rules that allow for fast new reductions and penalty calculations when needed. The method also allows us to make multiple assignments of variables. Numerical tests run on small tests show that the MVM over performs the original one in all instances while requiring comparable computing times. The tests also support the intuition that the new method provides optimal solutions almost all the time, making it a viable alternative to the classical transportation simplex.

1 Linear Transportation Problem

The linear transportation problem (LTP) consists in shipping a commodity from supply centers, called sources, to receiving centers, called destinations, while minimizing the total distribution cost. Assuming that we have m sources $i = 1, \cdots, m$

P. M. Takouda (✉)
School of Commerce & Administration, Laurentian University, Sudbury, Canada
e-mail: mtakouda@laurentian.ca

D. Almaatani
Department of Mathematics and Computer Sciences,
Laurentian University, Sudbury, Canada
e-mail: dalmaatani@laurentian.ca

Y. Gningue
Department of Mathematics and Computer Sciences, Laurentian University,
Sudbury, ON, Canada
e-mail: ygningue@cs.laurentian.ca

S. G. Diagne
Département de Mathématiques, Université Cheikh Anta Diop, Dakar, Sénégal
e-mail: gueyesalli@yahoo.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_3

and n destinations $j = 1, \cdots, n$, we denote by C_{ij} the cost of shipping one unit of commodity from source i to destination j, by a_i the capacity of source i and by b_j the demand at destination j, Then, the LTP can be formulated as follows.

$$\textbf{LTP} \begin{cases} \min CT = \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} X_{ij} \\ \sum_{j=1}^{n} X_{ij} = a_i \ ; \qquad i = 1, \cdots m \\ \sum_{i=1}^{m} X_{ij} = b_j \ ; \qquad j = 1, \cdots n \\ X_{ij} \geq 0 \ ; \quad i = 1, \cdots m \ ; \quad j = 1, \cdots n \end{cases} \tag{1}$$

It is a linear program with $n+m$ constraints and $n \times m$ variables, X_{ij} representing the quantity shipped from source i to destination j. The costs $C_{i,j}$ form a matrix $C = (C_{i,j})_{i,j}$ called the cost matrix of the problem.

LPT is usually optimally solved using the transportation simplex algorithm (see [3]). This algorithm has to be provided a starting basic feasible solution. Several methods exist that compute such starting points. One of the most efficient is the Vogel Approximation Method (VAM) [6]. Indeed, VAM produces good near optimal solutions, which reduce the number of iterations that the transportation simplex has to perform. Assuming for the rest of this chapter that a line in a matrix refers to either a row and a column in the matrix, the VAM runs as follows. For each line (row or column) of the cost matrix, compute its penalty, which is the *difference between the two least costs* of the line. Then, locate the line with the largest penalty (called the *penalty line*), and in that line, look for the lowest shipping cost (this cost is at the intersection of the penalty line and another one called the *complementary line*). Assign to the corresponding variable X the maximum quantity of commodity that can be shipped at that cost. Update corresponding demand and supply informations. One of these informations will be updated to 0. The corresponding line is said to be *saturated* and is removed from the cost matrix (which is shrinked). The process is repeated until all the demands have been satisfied. Some additional rules exist that helps deal with special situations such as the occurrence of several largest penalties or degeneracy (when the two penalty and complementary lines are saturated at the same time before the end of the algorithm). The interested reader should refer [6].

Modifications of the original VAM have been proposed, most of them for unbalanced transportation problems. For balanced ones, one variant proposed consisting in modifying the cost matrix as follows: from the cost matrix, obtain the row (respectively column) opportunity cost matrix by subtracting in each row (respectively column) the smallest cost of the row (column) from all entries in the row (column); then add the two row and column opportunity cost matrices to obtain a new cost matrix on which the original VAM is applied. In addition, some additional tie-breaking rules are proposed. These variants are proposed and analyzed in [5] and [7]. Our modification of the VAM extends the one proposed in [4] and is called the Modified Vogel Method (MVM).

The rest of the chapter is organized as follows. Next section presents the MVM. Then the method is illustrated on an example with four sources and five destinations in Sect. 3. Numerical tests are presented in Sect. 4 followed by concluding remarks in Sect. 5.

2 The Modified Vogel Method

The MVM can be described as follows. First, compute the reduced cost matrix, R, by applying successively a row and a column reduction on the cost matrix C and define a reduced transportation problem (RTP) by replacing C in LTP by R. Then, apply the VAM to RTP. At each iteration, the new shrunk matrix is reduced if needed.

A reduced cost matrix has a zero in each line. Therefore, penalties in MVM are simply the second lowest costs of the line. Due to the row and column reductions, each entry of the reduced cost matrix already contains information about gaps between the original costs in each row and column. Hence, the associated Vogel penalties are qualitatively better than the ones in VAM. Note also that using MVM, we compute solutions where some of the assigned variables are associated with zero reduced costs of the LTP. This makes them particularly appealing for the simplex transportation algorithms [4]. In fact, in several instances, the transportation simplex algorithm [2] is not needed: MVM is guaranteed to provide the optimal solutions.

Theorem 1 *The reduced transportation problem (RTP) is equivalent to the linear transportation problem (LTP), and if its optimal cost is zero, then the optimal solution of RTP is optimal for LTP.*

Proof The row and column reductions that have been applied to the cost matrix to obtain the reduced cost matrix are admissible transformations as defined in [1].

At each iteration of MVM, one line of the current reduced matrix is removed. The remaining shrunk matrix may not be in a reduced form. However, if the highest penalty is nonzero, the penalty line is saturated, and the penalty of the complementary line is zero, then, the shrunk cost matrix remains reduced. Indeed, all the lines parallel to the penalty line are unaltered. They stay reduced, and their penalties are unchanged. Then, the highest penalty being nonzero, there was only one zero entry on the penalty and that zero is also on the complementary line. Crossing the penalty do not remove a zero on the lines parallel to the complementary line. Hence, they stay reduced and their penalty would change only if their penalty was on crossed line. In such a case, the new penalty is simply the next smallest nonzero cost. Finally, the complementary line, since its penalty is 0, had at least two zero entries. Therefore, it has at least one zero remaining and stays reduced and its penalty has to be recalculated. As a result, only a few penalties have to be recalculated.

If the penalty of the complementary line was not zero, all the previous remarks are still valid except for the complementary line which is now not reduced. Again, it is easy to reduce: subtract its penalty from all the entries. Note that such an operation is equivalent to applying an admissible transformation (see [1]) to the reduced cost matrix to solve an equivalent problem in which the complementary line has a zero penalty. In summary, when the saturated line is the penalty line, the shrunk matrix is always reduced, up to an admissible transformation. Hence, the following results hold.

Theorem 2 *During the application of MVM, we have the two following assertions.*

1. *If no new reduction is necessary during the iterations, the solution obtained is optimal for LTP.*
2. *If all the successive line removals are associated to a unique largest penalty, then LTP is optimal.*

Remark 1 The only time we require the use of the transportation algorithm to obtain the optimal solution is when we have ties for the largest penalty and we had to do new matrix reduction. In the MVM algorithm, we keep track of the occurrences of tied largest penalties and new reductions.

The MVM algorithm is hence described as follows.

Modified Vogel Algorithm

```
Step 0.  Compute the reduced cost matrix (R).
    Set $Nred := 1$; $TieLpen := 0$ and $UniqueLpen := 1$.
Step 1.  Penalty Determination
    Determine the penalties $p_i$ for each row $i$ and $q_j$ for each column $j$.
    Find the largest penalty    $\max\{p_i, q_j\} = Lpen$.
    If there is a tie for the largest penalty then set    $TieLpen := 1$.
Step 2.  Assigning Variable
    Find a zero reduced cost in the line of $Lpen$: $R_{kr}$. $X_{kr}$ will be assigned a value.
Step 3.  Updating: assign the value of $X_{kr}$, updates $a_k$ and $b_r$.
    Eliminate the saturated line (supply or demand fully satisfied)
Step 4.  Stopping Test    If there is one remaining line then fill it and go to step 6
Step 5.  New Reduction of Remaining Matrix
    Reduce the remaining matrix then set $Nred := Nred + 1$.
    If    $TieLpen := 1$ then set $UniqueLpen := 0$ and go to step 1.
Step 6.  Otpimality Test
    If $Nred := 1$ then the MVM solution is optimal.
    elseif $UniqueLpen := 1$ the MVM solution is optimal.
        else
        find the dual variables and test the optimality.
```

3 Illustrative Example

Let’s illustrate the MVM on the following transportation problem.

1	3	8	7	6	**15**
8	10	7	9	6	**25**
2	11	6	8	9	**10**
5	9	8	7	7	**16**
10	**15**	**12**	**14**	**15**	

In each row, we identify the smallest cost and subtract it from all the row’s entries. As a result, there is a zero in each row. We say that the row is reduced. We obtain the following matrix.

0	2	7	6	5	**15**
2	4	1	3	0	**25**
0	91	4	6	7	**10**
0	4	3	2	2	**16**
10	**15**	**12**	**14**	**15**	

Since columns 1 and 5 already have a zero, they are already reduced. We reduce columns 2, 3, and 4 (again by subtracting the least cost in the entry from all the column’s entries). Then, we compute the penalties for each row (column 7) and column (row 6) and we start the MVM.

0	0	6	4	5	15	$p_1 = 0$	$p_1 = 4$	———
2	2	0	1	0	25	$p_2 = 0$	$p_2 = 0$	$p_2 = 0$
0	7	3	4	7	10	$p_3 = 3$	———	———
0	2	2	0	2	16	$p_4 = 0$	$p_4 = 2$	$p_4 = 2$
10	15	12	14	15				
$q_1 = 0$	$q_2 = 2$	$q_3 = 2$	$q_4 = 1$	$q_5 = 2$				
———	$q_2 = 2$	$q_3 = 2$	$q_4 = 1$	$q_5 = 2$				
———	———	$q_3 = 2$	$q_4 = 1$	$q_5 = 2$				
———	———	$q_3 = 2$	$q_4 = 1$	$q_5 = 2$				
———	———	$q_3 = 2$	$q_4 = 1$	———				

Iteration 1 $Lpen = 3$ leading to $X_{3,1} = 10$. No ties for $Lpen$. Both lines are saturated. Remove the line of largest penalty (row 3). Then, using the least initial costs in column 1, assign $X_{11} = 0$ and remove column 1. No new reduction needed.

Iteration 2 $Lpen = 4$ leading to $X_{12} = 15$. No ties for $Lpen$. Both lines are saturated. Remove the line of largest penalty (row 1). Again, the least initial cost in column 2 leads to $X_{41} = 0$ and remove column 2. No new reduction needed.

Iteration 3 $Lpen = 2$, and we have a three way tie: row 4, columns 3, 5. Lowest cost break the tie: assign $X_{25} = 15$, remove column 5. No new reduction.

Iteration 4 We assign X_{44}=14 and fill the last line. We obtain the optimal solution: X_{11}=0; X_{12}=15; X_{23}=10; X_{25}=15; X_{31}=10; X_{44}=14; X_{43}=2, with the total cost equal to $TC = 339$.

The solution is optimal for the LPT since no new reduction was ever required.

Note that in general, we will need to test the optimality by evaluating the dual variables and the reduced cost for the nonbasic variables.

4 Numerical Tests

We run some tests to compare VAM and MVM. Both codes were written in Java. We solved two sets of randomly generated LTPs. The first set consists of 12 problems having the same number (5, 10, 15) of sources and destinations. The second set contains 15 transportation problems with different number of sources and destinations. There were three to five sources and destinations.

The MVM has outperformed the VAM in all the cases. The improvement rate, measuring by how much MVM has improved the solution obtained by VAM, ranged from 0 to 20.92 %. At the same time, MVM has required less computing time: 11.17 s in average for VAM, compared to 9.72 s for MVM.

In all these problems, the MVM provides a better solution than the VAM. We noticed also that the MVM solution is optimal when it equals the VAM. This confirms results from the literature [5,7]: VAM provides an optimal solution at least 80 % of the times [5,7]. Our test suggests that MVM should provide the optimal solutions in a higher proportion of time. It is our next objective: compare the MVM's solutions with LTP optimal ones.

5 Conclusion

We introduced with MVM a new algorithm to compute approximate solutions to LTPs based on VAM. MVM always outperforms VAM without requiring significantly more time. It provides qualitatively better starting points to the transportation algorithms. It is proven to provide optimal solutions in several cases, and this optimality can be checked directly in the MVM algorithm. In the future, we would like to identify more of, if not all, the cases where we have a guarantee that MVM provides optimal solutions. It would allow MVM to become a viable alternative to the transportation simplex.

References

1. Burkard, R.E.: Admissible transformations and assignment problems. Vietnam J. Math. **35**(4), 373–386 (2007)
2. Charnes, A., Cooper, W.W.: The stepping-stone method for explaining linear programming calculations in transportation problems. Manage. Sci. **1**(1), 49–69 (1954)
3. Dantzig, G.B.: Linear Programming and Extensions. Princeton University Press (1963)
4. Diagne, S.G., Gningue, Y.: Méthode de Vogel modifiée pour la résolution des problèmes de transport simple. Appl. Math. Sci. **5**(48), 2373–2388 (2011)
5. Mathirajan, M., Neenakshi, B.: Experimental analysis of some variants of Vogel's Approximation Method. Asia Pac. J. Oper. Res. **21**(4), 447–462 (2004)
6. Reinfeld, N.V., Vogel, W.R.: Mathematical Programming. Prentice-Hall, Englewood Cliffs (1958)
7. Singh, S., Dubey, G.C., Shrivastava, R.: Optimization and analysis of some variants through Vogel's Approximation Method (VAM). IOSR J. Eng. (IOSRJEN) **2**(9), 20–30 (2012)

Input-to-State Stability of Large-Scale Stochastic Impulsive Systems with Time Delay and Application to Control Systems

M. S. Alwan, X. Z. Liu and W.-C. Xie

Abstract This chapter deals with large-scale nonlinear delay stochastic systems where the system states are subject to impulsive effects and perturbed by some disturbance input having bounded energy. The interest is to develop a comparison principle and establish input-to-state stability (ISS) in the mean square (m.s.) using vector Lyapunov function and Razumikhin technique. Impulses are being viewed as perturbation to stable systems, and they have a stabilizing role to unstable systems.

1 Introduction

Technology has been producing a new generation of high-dimensional, structurally sophisticated dynamical systems, known as *large-scale systems*. Typically, a large-scale system is described by a large number of variables, nonlinearities, and uncertainties. Nowadays, large-scale systems, as a tool, have been used to model numerous processes in many fields in science and engineering, such as large electric power network systems, control systems, aerospace systems, solar systems, nuclear reactors, chemistry, biology, and ecology systems. Readers may consult [5, 8].

A large class of systems in natural science and engineering are subjected to state changes over short time periods. The durations of these changes are often negligible when compared to the duration of the system process, so that these changes can be approximated as instantaneous changes of states or *impulses*. The resulting systems are called *impulsive systems* [4].

If time delay and random noise are considered in the later systems, we are led to *stochastic impulsive systems with time delay* [1, 2].

M. S. Alwan (✉) · X. Liu · W.-C. Xie
University of Waterloo, Waterloo, ON, Canada
e-mail: malwan@uwaterloo.ca

X. Liu
e-mail: xzliu@uwaterloo.ca

W.-C. Xie
e-mail: xie@uwaterloo.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_4

Input-to-state stability (ISS) is essential in modern nonlinear feedback and control system design. Generally, ISS studies the response of the forced system to a disturbance input where the underlying unforced system is asymptotically stable [3, 6, 7].

2 Problem formulation

Denote by $\mathbb{N}$ the set of natural numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, $\mathbb{R}^n$ the n-dimensional real space with the Euclidean norm $\|\cdot\|$, and $\mathbb{R}^{n\times m}$ the set of $n \times m$ matrices. If $g \in \mathbb{R}^{n\times m}$, its induced norm is $\|g\| = \sqrt{\text{trace}(g^T g)}$. Let $r > 0$ be the time delay, $\mathbb{C}([-r,0],\mathbb{R}^n)$ ($\mathbb{PC}([-r,0],\mathbb{R}^n)$) be space of continuous (piecewise continuous) functions ϕ mapping $[-r,0]$ into $\mathbb{R}^n$. If x is a function from $[t-r,\infty)$ to $\mathbb{R}^n$, then $x_t = x(t+s)$ for $s \in [-r,0]$ mapping $[-r,0]$ into $\mathbb{R}^n$, and $\|x_t\|_r = \sup_{t-r\le\theta\le t}\|x(\theta)\|$. Define $x_{t^-} \in \mathbb{PC}([-r,0],\mathbb{R}^n)$ by $x_{t^-}(s) = x(t+s)$ for $s \in [-r,0]$ and $x_{t^-}(s) = x(t^-)$ for $s = 0$. Let $W(t,\omega)$ denote an m-dimensional Wiener process.

Typically, an interconnected system with decomposition $\mathbb{D}_i$ may have the form

$$\mathbb{D}_i : \begin{cases} dw^i(t) = f_i(t,w_t^i)dt + g_i(t,w_t^1,w_t^2,\cdots,w_t^l)dt \\ \qquad\qquad + \sum_{j=1}^{l}\sigma_{ij}(t,w_t^j)dW_j(t), & t \neq \tau_k, \\ \Delta w^i(t) = \mathcal{I}_i(t,w_{t^-}^i), & t = \tau_k, \\ w_{t_0}^i = \phi_i(s), \qquad s \in [-r,0], \end{cases} \tag{1}$$

where $k \in \mathbb{N}$ and $i = 1,2,\cdots l$ for some $l \in \mathbb{N}$. w^i (or w_t^i) $\in \mathbb{R}^{n_i}$ is an n_i-dimensional vector state (or deviated state) and $n = \sum_i^l n_i$ for some $n_i \in \mathbb{N}$. $f_i : \mathbb{R}_+ \times \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$, $g_i : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n_i}$, $\sigma_{ij} : \mathbb{R}_+ \times \mathbb{R}^{n_j} \to \mathbb{R}^{n_i\times m_j}$, $m = \sum_i^l m_i$ for some $m_i \in \mathbb{N}$, $\mathcal{I}_i : \mathbb{T} \times \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ with $\mathbb{T} = \{\tau_k | k = 1,2,\cdots\}$ with impulsive moments $0 < \tau_1 < \tau_2 < \cdots$, and $\lim_{k\to\infty}\tau_k = \infty$, and $\phi_i : [-r,0] \to \mathbb{R}^{n_i}$. Define the isolated subsystems $\mathbb{S}_i$ by

$$\mathbb{S}_i : \begin{cases} dw^i(t) = f_i(t,w_t^i)dt + \sigma_{ii}(t,w_t^i)dW_i(t), & t \neq \tau_k, \\ \Delta w^i(t) = \mathcal{I}_i(t,w_{t^-}^i), & t = \tau_k, \\ w_{t_0}^i = \phi_i(s), s \in [-r,0]. \end{cases} \tag{2}$$

For $x \in \mathbb{R}^n$, let $x^T = [(w^1)^T,(w^2)^T,\cdots,(w^l)^T]$, and define $f : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ by $f^T(t,x_t) = [f_1^T(t,w_t^1), f_2^T(t,w_t^2),\cdots,f_l^T(t,w_t^l)]$, $g : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ by $g^T(t,x_t) = [g_1^T(t,x_t), g_2^T(t,x_t),\cdots,g_l^T(t,x_t)]$, $\sigma : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ by $\sigma(t,x_t) = [\sigma_{ij}(t,w_t^j)]$, $W : \mathbb{R}_+ \to \mathbb{R}^m$ by $W^T = [W_1^T, W_2^T,\cdots,W_l^T]$, where $W_i : \mathbb{R}_+ \to \mathbb{R}^{m_i}$, and impulsive functional $\mathcal{I} : \mathbb{T} \times \mathbb{R}^n \to \mathbb{R}^n$ by $\mathcal{I}^T(t,x_{t^-}) = [\mathcal{I}_1^T(t,w_{t^-}^1),\mathcal{I}_2^T(t,w_{t^-}^2),\cdots,\mathcal{I}_l^T(t,w_{t^-}^l)]$.

Then, the composite (or interconnected) system can be written in the form $\mathbb{S}$

$$\mathbb{S}:\begin{cases} dx(t) = F(t,x_t)dt + \sigma(t,x_t)dW(t), & t \neq \tau_k, \\ \Delta x(t) = \mathcal{I}(t,x_{t^-}), & t = \tau_k, \\ x_{t_0} = \Phi(s), \qquad s \in [-r,0], & \end{cases} \tag{3}$$

where $F(t,x_t) = f(t,x_t) + g(t,x_t)$, and $\Phi^T = [\phi_1^T, \phi_2^T, \cdots, \phi_l^T]$ with $\mathbb{E}[\|\Phi\|^2] < \infty$.

Definition 1 A function $\alpha \in \mathbb{C}(\mathbb{R}_+;\mathbb{R}_+)$ is said to belong to $\mathcal{K}$ (briefly, $\alpha \in \mathcal{K}$) if $\alpha(0) = 0$ and it is strictly increasing; it is said to belong to $\mathcal{K}_1$ (or $\mathcal{K}_2$) if $\alpha \in \mathcal{K}$ and it is convex (or concave). A function $\beta \in \mathbb{C}([0,a) \times \mathbb{R}_+;\mathbb{R}_+)$ is said to belong to class $\mathcal{KL}$ if, for each fixed s, the mapping $\beta(\cdot,s) \in \mathcal{K}$, and, for each fixed r, the mapping $\beta(r,\cdot)$ is decreasing and $\beta(r,s) \to 0$ as $s \to \infty$.

Definition 2 System (3) is said to be ISS in mean square (m.s.) if there exist functions $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that, for any x_{t_0} and bounded input u, the solution x satisfies

$$\mathbb{E}[\|x(t)\|^2] \leq \beta(\mathbb{E}[\|x_{t_0}\|_r^2], t - t_0) + \gamma\Big(\sup_{t_0 \leq \theta \leq t} \|u(\theta)\|\Big).$$

If, moreover, $\beta(\mathbb{E}[\|x_{t_0}\|_r^2], t - t_0) = K\mathbb{E}[\|x_{t_0}\|_r^2]e^{-\lambda(t-t_0)}$, for some positive constants K and λ, then system (3) is said to be exponential ISS in the m.s.

Definition 3 The isolated subsystem $\mathbb{S}_i$ in (2) is said to possess **Property A** if there exist functions $c_i \in \mathcal{K}_1$ and $a_i \in \mathbb{C}([\tau_{k-1},\tau_k) \times \mathbb{R}_+ \times \mathbb{R}^q;\mathbb{R})$, where $a_i(t,v,u)$ is concave in v for all $t \in \mathbb{R}_+$ and $u \in \mathbb{PC}(\mathbb{R}_+;\mathbb{R}^q)$, and $\lim_{(t,y,v)\to(\tau_k^-,x,u)} a_i(t,y,v) = a_i(\tau_k^-,x,u)$, and $V^i \in \mathbb{C}^{1,2}([-r,\infty) \times \mathbb{R}^n;\mathbb{R}_+)$, which is decrescent and satisfies

(i) $\forall (t,\psi^i(0)) \in [-r,\infty) \times \mathbb{R}^n$, $c_i(\|\psi^i(0)\|^2) \leq V^i(t,\psi^i(0))$, (a.s.), and, $\forall t \neq \tau_k$, $\psi^i \in \mathbb{PC}([-r,0];\mathbb{R}^n)$, and $u \in \mathbb{PC}(\mathbb{R}_+;\mathbb{R}^q)$,

$$\mathcal{L}_i V^i(t,\psi^i,u) \leq a_i(t, V^i(t,\psi^i(0)), u(t)), \quad \text{(a.s.)},$$

provided that $V^i(t+s,\psi^i(s)) \leq \bar{q}V(t,\psi^i(0))$ for some $\bar{q} > 1$ and $s \in [-r,0]$;

(ii) for any $\tau_k \in \mathbb{T}$ and $\psi^i \in \mathbb{PC}([-r,0];\mathbb{R}^n)$,

$$V^i(\tau_k, \psi^i(0) + \mathcal{I}_i(\tau_k, \psi^i(\tau_k^-))) \leq \alpha(d_k)V^i(\tau_k^-, \psi^i(0)), \text{(a.s.)},$$

where $\psi^i(0^-) = \psi^i(0)$ and $\prod_{k=1}^{\infty} \alpha(d_k) < \infty$ with $\alpha(d_k) > 1$ for all k.

3 Main results

Theorem 1 Comparison principle. *Assume that the following assumptions hold:*

(i) Every isolated subsystem $\mathbb{S}_i$ has Property A;

(ii) For any $i = 1, 2, \cdots, l$, there exist a function $\bar{b}_i \in \mathbb{C}([\tau_{k-1}, \tau_k) \times \mathbb{R}_+ \times \mathbb{R}^q; \mathbb{R})$ and $\bar{b}_i$ is quasi monotone nondecreasing such that

$$g_i^T(t, \psi, u)V^i_{\psi^i(0)}(t, \psi^i(0)) + \frac{1}{2} \sum_{j=1, i \neq j}^{l} tr[\sigma_{ij}^T(t, \psi^j, u)$$

$$\times V^i_{\psi^i(0)\psi^i(0)}(t, \psi^i(0))\sigma_{ij}(t, \psi^j, u)] < \bar{b}_i(t, V(t, \psi(0)), u),$$

where $V^T(t, x) = (V^1(t, w^1), \cdots, V^l(t, w^l))$;

(iii) Let $a^T(\cdot) = (a_1(\cdot), a_2(\cdot), \cdots, a_l(\cdot))$ and $\bar{b}^T(\cdot) = (\bar{b}_1(\cdot), \bar{b}_2(\cdot), \cdots, \bar{b}_l(\cdot))$, where $a_i(\cdot)$ and $\bar{b}_i(\cdot)$ are defined in (i) and (ii), respectively, and assume that

$$|a(t, v', u') + \bar{b}(t, v', u')|^2 \leq h_1(t) + h_2(t)\kappa(\|v'\|^2),$$

$$|a(t, v', u') + \bar{b}(t, v', u') - a(t, v'', u'') - b(t, v'', u'')| \leq K(\|v' - v''\| + \|u' - u''\|),$$

where $t \in \mathbb{R}_+$, h_1 and h_2 are $\mathbb{PC}(\mathbb{R}_+, \mathbb{R}_+)$ functions, $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, increasing, concave function, v' and $v'' \in \mathbb{R}^l_+$, u' and $u'' \in \mathbb{R}^q$, and $K > 0$;

(iv) There exists a function $p : \mathbb{R}_+ \times \mathbb{R}^l \times \mathbb{R}^q \to \mathbb{R}$ such that

$$\sup_{V(t,x) \leq v} \sum_{i,j=1}^{l} \|\sigma_{ij}^T(t, \psi^j, u)V_{\psi^i(0)^i(t, \psi^i(0))}(t, \psi^i(0))\|^2 \leq p(t, v, u)$$

$$\leq h_2(t)\kappa(\|v\|^2) + \gamma(\|u\|).$$

Then, $V(t_0, x_0) < v_0$ implies that $V(t, x(t)) < v(t) = (v^1, \cdots, v^l)^T$, where

$$\begin{cases} dv = [a(t, v, u) + \bar{b}(t, v, u)]dt + \mathcal{V}dW(t), & \forall t \geq t_0, \quad t \neq \tau_k, \\ \Delta v(t) = \alpha_M(d_k)v(t^-), & t = \tau_k, \end{cases} \tag{4}$$

with $\mathcal{V} = [v_{ij}]_{l \times l}$, $\|\mathcal{V}\|^2 \leq p(t, v, u)$, and $\alpha_M(\cdot) = \max_i\{\alpha_i(\cdot)\}$.

Proof Define $V^T(t, x(t)) = (V^1(t, w^1), \cdots, V^l(t, w^l))$, where V^i is the Lyapunov function of ith subsystem. Then, $dV^T(t, x(t)) = (dV^1(t, w^1), \cdots, dV^l(t, w^l))$, where

$$dV^i(t, w^i) < [a_i(t, V^i(t, w^i), u) + b_i(t, V^i(t, w^i), u)]dt + \sum_{ij}^{l} y_{ij}dW_i(t),$$

with $y_{ij} = V^{i^T}_{w^i}(t, w^i)\sigma_{ij}(t, w^j_t, u)$. It follows that, for all $t \in [\tau_{k-1}, \tau_k)$, $k = 1, 2, \cdots$,

$$dV(t, x(t)) < [a(t, V(t, x(t)), u(t)) + b(t, V(t, x(t)), u(t))]dt + YdW(t).$$

At $t = \tau_k$, one can get $V^T(t, x(t)) \le \alpha_M(d_k)V^T(t^-, x(t^-))$. Particularly, for $t \in [\tau_0, \tau_1)$, we have $V^i(t_0, w^i(t_0)) < y_0$ and

$$dV^i(t, w^i) - dy_i < \big\{[a_i(t, V^i(t, w^i), u) - a_i(t, y_i, u)] + [b_i(t, V^i(t, w^i), u) - b_i(t, y_i, u)]\big\}dt.$$

By Theorem 4.5.2 in [5], $V^i(t, w^i(t)) < y_i(t)\ \forall t \in [\tau_0, \tau_1)$, and, at $t = \tau_1$, we have

$$V^i(\tau_1, w^i(\tau_1)) - y_i(\tau_1) < \alpha_M(d_k)\big[V^i(\tau_1^-, w^i(\tau_1^-)) - y_i(\tau_1^-)\big] < 0,$$

i.e., $V^i(\tau_1, w^i(\tau_1)) < y_i(\tau_1)$. Similarly, for $k = 1, 2, \cdots$ and $t \in [\tau_{k-1}, \tau_k)$, $V^i(t, w^i(t)) < y_i(t)$ and, at $t = \tau_k$, $V^i(\tau_k, w^i(\tau_k)) < y_i(\tau_k)$. Therefore, for all $t \ge t_0$, and $i = 1, 2, \cdots, l$, $V_i(t, w^i(t)) < y_i(t)$, which implies that $V(t, x(t)) < y(t)$, $\forall\, t \ge t_0$, as required.

Theorem 2 Stability results. *Suppose that the assumptions of Theorem 1 hold, and there exist $\alpha \in \mathcal{K}_2$, $c \in \mathcal{K}_1$, a function $\bar{h} \in \mathbb{C}([\tau_k, \tau_{k-1}) \times \mathbb{R}^l; \mathbb{R}_+)$, $z \in \mathbb{R}^l$, and $U \in \mathbb{C}^{1,2}([\tau_k, \tau_{k-1}) \times \mathbb{R}^l : \mathbb{R}_+)$ which is decrescent, $U(t, 0) = 0$, and satisfies*

(i) For all $t \in \mathbb{R}_+$ and $y \in \mathbb{PC}(\mathbb{R}_+; \mathbb{R}^l)$, $\alpha(\|y\|^2) \le U(t, y)$, $z^T U_{yy}(t, y)z \le \bar{h}(t, y)\|z\|^2$, and

$$U_t(t, y) + U_y(t, y)[a(t, y, u) + b(t, y, u)] + \frac{1}{2}h(t, y)p(t, y, u) \le -c(\|y\|)$$

whenever $\|y\| > V^i(t, w^i) \ge \rho(\|u\|)$ for some $\rho \in \mathcal{K}$ and i;
(ii) For any $\tau_k \in \mathbb{T}$ and $y \in \mathbb{PC}(\mathbb{R}_+; \mathbb{R}^l)$, $U(\tau_k, y(\tau_k)) = \alpha(d_k)U(\tau_k^-, y(\tau_k^-))$.

Then, comparison system (4), and hence composite system (3) are ISS in m.s.

Proof Let $y \ge 0$ be the solution of (4). Applying the Itô formula to U gives

$$\mathcal{L}U(t, y, u) \le -c(\|y\|), \quad \text{whenever } \|y\| \ge \rho(\|u\|).$$

By the previous analysis, (4) has the desired stability property. As for the composite system (3), we have shown in Theorem 1 that $V(t, x(t)) < y(t)$ holds for all $t \ge t_0$, and, from (i), we obtain $\|y\| > \|V(t, x)\| \ge V^i(t, w^i) \ge \rho(\|u\|)$. It follows that

$$c(\|x(t)\|^2) \le \Big[\sum_{i=1}^{l} c_i^2(\|w^i\|^2)\Big]^{1/2} \le \|V(t, x(t))\| < \|y(t)\|, \qquad c \in \mathcal{K}_1.$$

Taking the mathematical expectation and applying c^{-1} implies the desired result.

3.1 Application. Control system

Example 1 Consider the control system, which describes the longitudinal motion of an aircraft. This example is a modification of Example 4.6.1 in [5].

$$\begin{cases} \mathrm{d}x = Axdt + bf(y)\mathrm{d}t + \sigma_{11}(x(t-1))\mathrm{d}W_1 + \sigma_{12}(y)\mathrm{d}W_2, \qquad t \ne \tau_k, \\ \mathrm{d}y = (-\,\zeta y - \xi f(y) + u)\mathrm{d}t + a^T x\mathrm{d}t + \sigma_{21}(x)\mathrm{d}W_1 + \sigma_{22}(y(t-1))\mathrm{d}W_2, t \ne \tau_k, \end{cases} \tag{5}$$

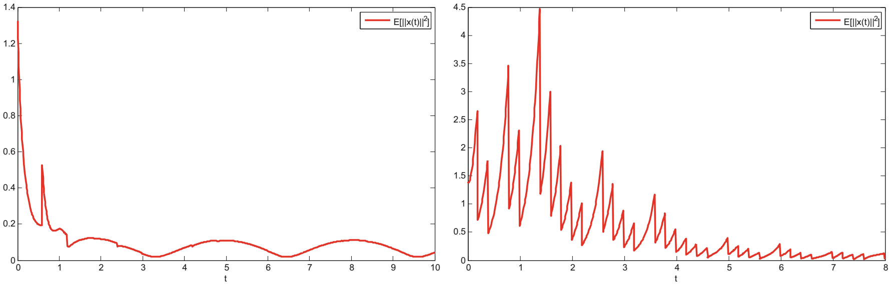


Fig. 1 Mean square input-to-state stability (*left*) and stabilization (*right*) of $(x^T, y)^T$ where $u(t) = \sin(t)$.

where $x^T = (x_1, x_2, x_3, x_4)$ is the system state, $y \in \mathbb{R}$ is the controller (i.e., $n_1 = 4$, $n_2 = 1$), $A \in \mathbb{R}^{4\times 4}$, $b \in \mathbb{R}^4$, $\zeta, \xi \in \mathbb{R}$, $f \in \mathbb{R}$ is continuous for all $y \in \mathbb{R}$, $f(y) = 0$ if and only if $y = 0$, and $0 < yf(y) < k|y|^2$ for all $y \neq 0$ and $k > 0$, $u \in \mathbb{R}$, $a \in \mathbb{R}^4$, $\sigma_{11} \in \mathbb{R}^{4\times 4}$, $\sigma_{12} \in \mathbb{R}^{1\times 1}$, $\sigma_{21} \in \mathbb{R}^{4\times 1}$, $\sigma_{22} \in \mathbb{R}^{1\times 1}$, $W_1 \in \mathbb{R}^4$, and $W_2 \in \mathbb{R}$. Let

$$A = \begin{pmatrix} -5 & 0 & 0 & 0 \\ 0 & -6 & 0 & 0 \\ 0 & 0 & -8 & 0 \\ 0 & 0 & 0 & -10 \end{pmatrix}, \sigma_{11} = 0.01$$

$$\begin{pmatrix} \sin x_1(t-1) & 0 & \frac{x_2(t-1)}{1+x_4^2} & 0 \\ 0 & \frac{x_2(t-1)}{1+x_1^2} & 0 & -x_3^2(t-1) \\ 0 & 0 & x_3(t-1) & 0 \\ 0 & 0 & 0 & -x_4(t-1) \end{pmatrix},$$

$b^T = (1, 1, 1, 1, 1)$, $a^T = (1, 1, 1, 1)$, $\zeta = 5$, $\xi = 2$, $\sigma_{12} = \frac{0.01y}{1+y^2}$, $\sigma_{21}^T = 0.01\,(x_2, x_1, x_4, x_3)$, $\sigma_{22} = 0.01 \sin y(t-1)$, and $u(t) = \sin(t)$. The impulses are given by

$$\begin{cases} \Delta x(\tau_k) = \mathcal{I}_1(\tau_k, x(\tau_k^-)) = \frac{1}{k^2}(-\,2x_1(\tau_k^-), -2x_2(\tau_k^-), 2x_3(\tau_k^-), 0)^T, \\ \Delta y(\tau_k) = \mathcal{I}_2(\tau_k, y(\tau_k^-)) = -\frac{1}{1+k^2} y(\tau_k^-). \end{cases} \tag{6}$$

Let $V^1(x) = \|x\|^2$ and $V^2(y) = y^2$. One can show the conditions are satisfied with $\tau_{k+1} - \tau_k \geq 0.6$ [2], i.e., $(x^T, y)^T \equiv (0^T, 0)$ is exponentially stable in the m.s. Applying the disturbance $u(t) = \sin(t)$, the composite system is ISS in m.s. See Fig. 1 (left).

Example 2 Reconsider the control composite continuous system (5) with *unstable state subsystem* in which the entry a_{11} of matrix A is changed to 5, and the impulsive difference equations are defined by $\Delta x(\tau_k) = -\frac{5}{4}x(\tau_k^-)$, $\Delta y(\tau_k) = -\frac{5}{4}y(\tau_k^-)$. Then, one gets $\tau_k - \tau_{k-1} < 0.33$ for all k. That is, the solution has been stabilized by the impulsive effects. See Fig. 1 (right).

Acknowledgment The research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Alwan, M.S., Liu, X.Z., Xie, W-C.: Existence, continuation, and uniqueness problems of stochastic impulsive systems with time delay. J. Frankl. Inst. **347**, 1317–1333 (2010)
2. Alwan, M.S.: Qualitative properties of stochastic hybrid systems and applications. Ph. D. Thesis, University of Waterloo, Ontario, Canada (2011)
3. Hespanha, J.P., Liberzon, D., Tell, A.R.: On input-to-state stability of impulsive systems. Proceeding of the 44th Conference on Decision and Control, and the European Control Conference, Seville, Spain, 12–15 December (2005)
4. Lakshmiknatham, V., Bainov, D.D., Simeonov, P.S.: Theory of Impulsive Differential Equations. World Scientific, Singapore (1989)
5. Michel, A.N., Miller, R.K.: Qualitative Analysis of Large Scale Dynamical Systems. Academic, New York (1977)
6. Sontag, E.D.: Smooth stabilization implies coprime factorization. IEEE Trans. Automat. Control **34**(4), 435–443 (1989)
7. Teel, A.R, Moreau, L, Nesic, D.: A unified framework for input-to-state stability in systems with two time scales. IEEE Trans. Automat. Control **48**(9), 1526–1544 (2003)
8. Zecevic, A., Silijak, D.D.: Control of Complex Systems: Structural Constraints and Uncertainty. Springer, New York (2010)

Replicator Dynamics of Axelrod's Norms Games

Michael Andrews, Edward Thommes and Monica G. Cojocaru

Abstract We create pure strategy versions of Robert Axelrod's well-known norms and metanorms games. Our findings show that the only evolutionarily stable strategy (ESS) in the norms game is one in which a player defects and is lenient. This result is derived using classic game theoretical tools, and we conclude that Axelrod's original statement that the norms game always collapses holds. The metanorms game, however, has two evolutionarily stable strategies. The first is a repeat from the norms game, while the other is one in which a player follows the norm and punishes those who are lenient and those who defect.

1 Introduction

In a social setting, a *norm* can be defined as an established set of rules or behaviors that individuals are expected to follow, and be punished for not following [2]. The concept of social norms has become an increasingly popular topic over the last two decades, given that they are an essential part of group living [3–5, 12]. Thus, studying norms and their establishment in a society may allow us to understand group behavior on a more fundamental level [8]. Axelrod first introduced a game-theoretic approach to the social sciences [1], and in his well cited paper [2], he constructs two n-player evolutionary games that seek to model the establishment of norms.

The goal of evolutionary game theory is to model the behavior of an evolving population of players from generation to generation. Much like classical game theory, an evolutionary approach involves a player employing a chosen strategy in some contest against one or more adversaries. The rules of this contest, or game, dictate

M. Andrews (✉) · M. G. Cojocaru
University of Guelph, 50 Stone Rd E, Guelph, ON, Canada
e-mail: mandre04@uoguelph.ca

M. G. Cojocaru
e-mail: mcojocar@uoguelph.ca

E. Thommes
GlaxoSmithKline, 7333 Mississauga Rd N, Mississauga, ON, Canada
e-mail: ethommes@uoguelph.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_5

$$\begin{array}{c|cccc}
 & NP & NL & DP & DL \\
NP & 0 & 0 & CE+H & CE+H \\
NL & 0 & 0 & H & H \\
DP & \frac{T}{N_{Pop}}+CP & \frac{T}{N_{Pop}} & \frac{T}{N_{Pop}}+CP+CE+H & \frac{T}{N_{Pop}}+CE+H \\
DL & \frac{T}{N_{Pop}}+CP & \frac{T}{N_{Pop}} & \frac{T}{N_{Pop}}+CP+H & \frac{T}{N_{Pop}}+H
\end{array}$$

the payoff, or fitness, each player receives when these strategies are played against one another.

Axelrod chose to analyze his norms games using agent-based model (ABM) simulations [2]. In his original game, the *norms game*, players in a population (of constant size N_{Pop}) can choose to defect, and also choose to punish those they have seen defecting. Players who defect not only receive a temptation payoff of 3 ($T = 3$) but also have a chance of being caught (C), which is chosen to be uniform between 0 and 1. We decide to use the expected value of this variable ($C = 0.5$) in our analysis. The players who are caught have a chance of being punished for a payoff of -9 ($P = -9$) by all of those who see them. However, each player that chooses to punish must pay an enforcement cost of -2 ($E = -2$). Players that do not defect (i.e., follow the norm) will get hurt by all those that do, receiving a payoff of -1 ($H = -1$) each time.

In this game, players choose to defect or punish based on their *boldness* (B) and *vengefulness* (V), respectively. A high boldness corresponds to a high probability of a player defecting, and a high vengefulness corresponds to a high probability of a player punishing another player they have seen defecting.

Recently, Axelrod's games have been subject to more rigorous testing, once again using the approach of ABM simulations. For examples of this, see Mahmoud et al. [9], and Galan and Izquierdo [6]. In our work, we recreate Axelrod's norm establishment games using a pure strategy analytical approach. We begin in Sect. 2 by utilizing these pure strategy mechanics to analyze Axelrod's norms game. Then, in Sect. 3, we take a similar approach with Axelrod's metanorms game. We then end in Sect. 4 with discussion and concluding remarks.

2 Evolutionary Norms Game

We can view Axelrod's norms game as one with four possible pure strategies to play. These are to follow the norm and punish (*NP*), follow the norm and be lenient (NL), defect and punish (DP), and defect and be lenient (DL). The payoff matrix corresponding to this game can be written as follows if we consider a scenario where two players from the population play against one another

We note that this game is symmetric. That is, all players have the same strategy set and payoffs. We also note that defectors receive a payoff of $\frac{T}{N_{Pop}}$. This reflects Axelrod's construction of his game, where instead of two players competing against each other, one player will play against the entire population. For example, in his simulations, it is possible for a defector to be punished by all other players at once. If this happens, the payoffs of this game dictate that the defector will obtain the full temptation payoff T, but also receive punishment from every other player, $(N_{Pop} - 1)P$. We choose to replace $N_{Pop} - 1$ with N_{Pop}, which will become a better approximation when N_{Pop} is large. In fact, this approximation, even with Axelrod's relatively small original population size of 20, does not significantly change our results.

2.1 Replicator Dynamics of the Norms Game

In this section, we analyze the matrix game above using the continuous replicator equations [15].

The payoffs corresponding to each of the four possible strategies are

$$\begin{aligned}
\pi_{NP} &= H(S_{DP} + S_{DL})N_{Pop} + CE(S_{DP} + S_{DL})N_{Pop} \qquad (1)\\
\pi_{NL} &= H(S_{DP} + S_{DL})N_{Pop}\\
\pi_{DP} &= T + CP(S_{NP} + S_{DP})N_{Pop} + CE(S_{DP} + S_{DL})N_{Pop} +\\
&\quad H(S_{DP} + S_{DL})N_{Pop}\\
\pi_{DL} &= T + CP(S_{NP} + S_{DP})N_{Pop} + H(S_{DP} + S_{DL})N_{Pop},
\end{aligned}$$

where S_{NP} is the fraction of the population that plays strategy NP, S_{NL} is the fraction of the population that plays strategy NL and so forth. Our differential equation system then looks as follows:

$$\begin{aligned}
\dot{S}_{NP} &= S_{NP}[\pi_{NP} - \bar{\pi}]\\
\dot{S}_{NL} &= S_{NL}[\pi_{NL} - \bar{\pi}]\\
\dot{S}_{DP} &= S_{DP}[\pi_{DP} - \bar{\pi}]\\
\dot{S}_{DL} &= S_{DL}[\pi_{DL} - \bar{\pi}] \qquad (2)
\end{aligned}$$

with $\bar{\pi}$ denoting the average payoff of the population in a given state. Also, we note that $S_{NP} + S_{NL} + S_{DP} + S_{DL} = 1$, with each fraction taking values between 0 and 1. In order to remain consistent with Axelrod [2], we wish to convert our strategies into the terms of *vengefulness* and *boldness*. Thus, we define $V = S_{NP} + S_{DP}$ and $B = S_{DP} + S_{DL}$. Moreover, we apply the fact that all fractions sum to 1 and obtain the relation

$$S_{NL} = 1 - V - B + S_{DP}.$$

Our differential equation system (3) transforms into

$$\dot{B} = S_{DP}BCEN_{Pop} + BT + BVCPN_{Pop} - B^2VCEN_{Pop} - B^2T - B^2VCPN_{Pop}$$

$$\dot{V} = VBCEN_{Pop} - V^2BCEN_{Pop} - VBT - V^2BCPN_{Pop} + S_{DP}T + S_{DP}VCPN_{Pop}$$
$$\dot{S}_{DP} = -S_{DP}(-T - CPVN_{Pop} - CEBN_{Pop} + VBCEN_{Pop} + BT + BVCPN_{Pop}), \tag{3}$$

with the constraints

$$\begin{aligned} 0 &\le B + V - S_{DP} \le 1 \\ 0 &\le B - S_{DP} \le 1 \\ 0 &\le V - S_{DP} \le 1 \\ 0 &\le B, V, S_{DP} \le 1 \end{aligned} \tag{4}$$

We find from using both analytic results and observing simulations of this system that the only locally asymptotically stable equilibrium point corresponds to $B = 1$, $V = 0, S_{DP} = 0$. This equilibrium directly corresponds to the population playing strategy DL. By the folk theorem of evolutionary game theory [7], we claim that this state is a Nash equilibrium [10, 11]. Moreover, by using the game's payoff matrix we observe that DL is a strict Nash equilibrium. Once again by the folk theorem, it is thus locally asymptotically stable under the replicator dynamics (3). We see that DL is an evolutionarily stable strategy (ESS) [13, 14].

Using the definition of an ESS along with our norms game payoff matrix and the folk theorem, we see that the population playing strategy DL is in fact evolutionarily stable. Thus, we have verified Axelrod's original claim which he gathered from his simulation results by using a classical game theory approach.

3 Evolutionary Metanorms Game

We now consider Axelrod's second game, his so-called *metanorms game*. In this version, players have similar strategies and payoffs as the norms game, but now also have the opportunity to punish a non-punisher, provided that they are seen not punishing. Axelrod makes the assumption that a player who will punish a defector will also punish a non-punisher. Similarly, a player who is lenient toward defectors will also be lenient toward non-punishers. We note that without these restrictions, the metanorms game ESS dynamics break down to be identical to that of the norms game.

The payoffs for the metanorms game are as follows. The punishment a player receives for being lenient toward defectors is denoted P', and the enforcement cost associated with punishing non-punishers is E'. Similar to Axelrod's research, we simply use $P' = P$ and $E' = E$. The symmetric game matrix for the metanorms game is then:

We note here that in the payoff matrices for both the norms and metanorms games, the addition of the payoff H in the DP and DL columns does not impact either the ESS structure or the replicator dynamics in any way. Thus, identical results would be obtained if H is removed entirely.

$$\begin{array}{c} \\ NP \\ NL \\ DP \\ DL \end{array} \begin{array}{c} \begin{array}{cccc} NP & NL & DP & DL \end{array} \\ \begin{pmatrix} 0 & CE' & CE+H & CE+CE'+H \\ CP' & 0 & CP'+H & H \\ \frac{T}{N_{Pop}}+CP & \frac{T}{N_{Pop}}+CE' & \frac{T}{N_{Pop}}+CP+CE+H & \frac{T}{N_{Pop}}+CE+CE'+H \\ \frac{T}{N_{Pop}}+CP+CP' & \frac{T}{N_{Pop}} & \frac{T}{N_{Pop}}+CP+CP'+H & \frac{T}{N_{Pop}}+H \end{pmatrix} \end{array} .$$

3.1 Replicator Dynamics of the Metanorms Game

Using the same process and change of variables as in the regular norms game, we obtain the same feasible region (4), and our differential equation system in terms of V, B, and S_{DP} becomes

$$\begin{aligned} \dot{B} =& S_{DP}CEN_{Pop} - VBCEN_{Pop} + S_{DP}BCEN_{Pop} - S_{DP}VCEN_{Pop} + BVCPN_{Pop} - \\ & S_{DP}VCPN_{Pop} + V^2BCPN_{Pop} + V^2BCEN_{Pop} - \\ & B^2VCEN_{Pop} - B^2VCPN_{Pop} + BT - B^2T \\ \dot{V} =& -CPV^2N_{Pop} + CPV^3N_{Pop} + CEV^3N_{Pop} + CEVN_{Pop} - 2CEV^2N_{Pop} - \\ & VBT + VBCEN_{Pop} + S_{DP}VCPN_{Pop} - V^2BCPN_{Pop} - \\ & V^2BCEN_{Pop} + S_{DP}T \\ \dot{S}_{DP} =& -S_{DP}(-CEN_{Pop} - T - CPV^2N_{Pop} - CEBN_{Pop} + 2CEVN_{Pop} - \\ & CEV^2N_{Pop} + VBCEN_{Pop} + BVCPN_{Pop} + BT) \end{aligned} \tag{5}$$

In this game, analytic results and simulations of the system show that bistability exists between two locally asymptotically stable equilibrium points. These correspond to points $B = 1, V = 0, S_{DP} = 0$ and $B = 0, V = 1, S_{DP} = 0$. By the folk theorem [7], we state that these strategies are Nash equilibria. Moreover, from the payoff matrix we can see that NP and DL are strict Nash equilibria and thus locally asymptotically stable under (3.1) and are therefore both ESS's.

Again, we bring quantitative evidence that supports Axelrod's original claims on the behavior of his metanorms game. In this case, we show that two evolutionarily stable states exist.

4 Discussion and Conclusions

We have reanalyzed Axelrod's norms and metanorms games using a pure strategy game theoretical approach. We have found from using this analysis that one ESS exists in the norms game corresponding to a complete norm collapse. We also note that an equilibrium in which the boldness of a population is 0, or there is a mix of strategies NL and NP being played, is unstable and will ultimately lead to a norm collapse over time.

In addition, Axelrod's metanorms game has two ESSs. The first is a repeat from the norms game, that is, a player defects and is always lenient. The other is one in which a player follows the norm and punishes those who are lenient and those who defect. Given certain initial conditions, the population of players can either evolve to a state in which the norm collapses, or to a state in which the norm is established.

References

1. Axelrod, R.: The Evolution of Cooperation. Basic Books, Member of Perseus Basic Books, Cambridge MA (1984)
2. Axelrod, R.: An evolutionary approach to norms. Am. Polit. Sci. Rev. **80**(4), 1095–1111 (1986)
3. Axtell, R., Epstein, J., Young, P.: The emergence of classes in a multiagent bargaining model. In Social Dynamics, pp. 191–211. MIT Press (1999)
4. Epstein, J.: Learning to be thoughtless: social norms and individual computation. Comput. Econ. **18**, 9–24, 2001.
5. Epstein, J., Axtell, R.: Growing Artificial Societies: Social Science From the Bottom Up. MIT Press (1996)
6. Galan, J., Izquierdo, L.: Appearances can be deceiving: lessons learned re-implementing Axelrod's 'evolutionary approach to norms'. J. Artif. Soc. Soc. Simul. **8**(3), 2 (2005). http://jasss.soc.surrey.ac.uk/8/3/2.html
7. Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press (1998)
8. Kameda, T., Takezawa, M., Hastie, R.: The logic of social sharing: an evolutionary game analysis of adaptive norm development. Personal. Soc. Psychol. Rev. **7**(1), 2–19 (2003)
9. Mahmoud, S., Griffiths, N., Keppens, J., Luck, M.: An analysis of norm emergence in Axelrod's model. European Workshop on Multi-Agent Systems (2010)
10. Nash, J.: Equilibrium points in n-person games. Proc. Natl. Acad. Sci. **36**, 48–49 (1950)
11. Osborne, M., Rubinstein, A.: A Course in Game Theory. MIT Press (1994)
12. Schelling, T.: Micromotives and Macrobehavior. Norton (1978)
13. Smith, J.M.. Evolution and the Theory of Games. Cambridge University Press (1982)
14. Smith, J.M., Price, G.: The logic of animal conflict. Nature **246**, 15–18 (1973)
15. Taylor, P., Jonker, L.. Evolutionarily stable strategies and game dynamics. Math. Biosci. **40**, 145–156 (1978)

Computing Least Squares Condition Numbers on Hybrid Multicore/GPU Systems

M. Baboulin, J. Dongarra and R. Lacroix

Abstract This chapter presents an efficient computation for least squares conditioning or estimates of it. We propose performance results using new routines on top of the multicore-GPU library MAGMA. This set of routines is based on an efficient computation of the variance–covariance matrix for which, to our knowledge, there is no implementation in current public domain libraries LAPACK and ScaLAPACK.

1 Introduction

Linear least squares (LLS) is a classical linear algebra problem in scientific computing, arising for instance in many parameter estimation problems [5]. We consider the overdetermined full rank linear least squares problem $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$, with $A \in \mathbb{R}^{m \times n}, m \geq n$ and $b \in \mathbb{R}^m$.

In addition to computing LLS solutions efficiently, an important issue is to assess the numerical quality of the computed solution. The notion of conditioning provides a theoretical framework that can be used to measure the numerical sensitivity of a problem solution to perturbations. Similarly to [2, 3], we suppose that the perturbations on data are measured using the Frobenius norms for matrices and the Euclidean norm for vectors. Then we can derive simple formulas for the condition number of the LLS solution x or its components using the R factor (from the QR decomposition of A), the residual and x. We can also use the variance–covariance matrix.

M. Baboulin (✉)
Inria and University of Paris-Sud, Orsay, France
e-mail: marc.baboulin@inria.fr

J. Dongarra
University of Tennessee, Knoxville, TN, USA
e-mail: dongarra@eecs.utk.edu

R. Lacroix
Inria and University Pierre et Marie Curie, Paris, France
e-mail: remi.lacroix@inria.fr

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_6

In this chapter, we propose algorithms to compute LLS condition numbers in a computational time that is affordable for large-scale simulations, in particular using the variance–covariance matrix. We also compute statistical condition estimates that can be obtained cheaply ($\mathcal{O}(n^2)$ operations) and with a satisfying accuracy using an approach similar to [6, 8]. For these algorithms, we describe an implementation for LLS conditioning using the MAGMA library [4, 10], which is a dense linear algebra library for heterogeneous multicore-GPU architectures with interface similar to LAPACK. Our implementation takes advantage of current hybrid multicore-GPU systems by splitting the computational work between the GPU and the multicore host. We present performance results, and these results are compared with the computational cost for computing the LLS solution itself.

2 Closed Formulas and Statistical Estimates

In this section, we recall some existing formulas to compute or estimate the condition number of an LLS solution x or of its components. We suppose that the LLS problem has already been solved using a QR factorization (the normal equations method is also possible but the condition number is then proportional to $cond(A)^2$). Then the solution x, the residual $r = b - Ax$, and the factor $R \in \mathbb{R}^{n \times n}$ of the QR factorization of A are readily available.

From [3], we obtain a closed formula for the absolute condition number of the LLS solution as

$$\kappa_{LS} = \|R^{-1}\|_2 \left(\|R^{-1}\|_2^2 \|r\|_2^2 + \|x\|_2^2 + 1\right)^{\frac{1}{2}}, \tag{1}$$

where x, r and R are exact quantities.

We can also compute $\overline{\kappa}_{LS}$, statistical estimate of κ_{LS} that is obtained using the condition numbers of $z_i^T x$ where $z_1, z_2, ..., z_q$ are q random orthogonal vectors of $\mathbb{R}^n$, obtained for instance via a QR factorization of a random matrix $Z \in \mathbb{R}^{n \times q}$. The condition number of $z_i^T x$ can be computed using the expression given in [3] as

$$\kappa_i = \left(\|R^{-1}R^{-T}z_i\|_2^2 \|r\|_2^2 + \|R^{-T}z_i\|_2^2 (\|x\|_2^2 + 1)\right)^{\frac{1}{2}}. \tag{2}$$

Then $\bar{\kappa}_{LS}$ is computed using the expression $\bar{\kappa}_{LS} = \frac{\omega_q}{\omega_n}\sqrt{\sum_{j=1}^{q} \kappa_j^2}$ with $\omega_q = \sqrt{\frac{2}{\pi(q-\frac{1}{2})}}$. As explained in [6], choosing $q = 2$ random vectors enables us to obtain a satisfying accuracy.

By considering in Eq. (2) the special case where $z_i = e_i$ where e_i is a canonical vector of $\mathbb{R}^n$, we can express the condition number of the component $x_i = e_i^T x$ in Eq. (3). Then we can calculate a vector $\kappa_{CW} \in \mathbb{R}^n$ with components κ_i being the exact condition number of x_i and expressed by

$$\kappa_i = \left(\|R^{-1}R^{-T}e_i\|_2^2 \|r\|_2^2 + \|R^{-T}e_i\|_2^2 (\|x\|_2^2 + 1)\right)^{\frac{1}{2}}. \tag{3}$$

We can also find in [6, 8] a statistical estimate for each κ_i.

3 Variance–Covariance Matrix

In many physical applications, LLS problems are expressed using a statistical model often referred to as *linear statistical model* where we have to solve

$$b = Ax + \epsilon,\ A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m,$$

with ϵ being a vector of random errors having expected value 0 and variance-covariance $\sigma_b^2 I$. The matrix A is called the regression matrix and the unknown vector x is called the vector of regression coefficients. Following the Gauss–Markov theorem [11], the least squares estimate $\hat{x}$ is the linear unbiased estimator of x satisfying $\hat{x} = arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$,

with minimum variance–covariance equal to

$C = \sigma_b^2 (A^T A)^{-1}$.

The diagonal elements c_{ii} of C give the variance of each component $\hat{x}_i$. The off-diagonal elements $c_{ij},\ i \neq j$ give the covariance between $\hat{x}_i$ and $\hat{x}_j$. Then instead of computing condition numbers (which are notions more commonly handled by numerical linear algebra practitioners) physicists often compute the variance-covariance matrix whose entries are intimately correlated with condition numbers κ_i and κ_{LS} mentioned previously.

When the variance–covariance matrix has been computed, the condition numbers can be easily obtained. Indeed, we can use the fact that $\left\|R^{-1}\right\|_2^2 = \frac{\|C\|_2}{\sigma_b^2}$, $\|R^{-T}e_i\|_2^2 = \frac{c_{ii}}{\sigma_b^2}$, and $\|R^{-1}R^{-T}e_i\|_2 = \frac{\|C_i\|_2}{\sigma_b^2}$ where C_i and c_{ii} are respectively the ith column and the ith diagonal element of the matrix C. Then by replacing respectively in Eqs. (1) and (3), we get the formulas

$$\kappa_{LS} = \frac{\|C\|_2^{1/2}}{\boldsymbol{\sigma}_b}((m-n)\|C\|_2 + \|x\|_2^2 + 1)^{1/2}, \tag{4}$$

and

$$\kappa_i = \frac{1}{\boldsymbol{\sigma}_b}((m-n)\|C_i\|_2^2 + c_{ii}(\|x\|_2^2 + 1))^{1/2}. \tag{5}$$

Note that, when $m > n$, $\frac{1}{m-n}\ \|r\|_2^2$ is an unbiased estimate of $\boldsymbol{\sigma}_b^2$ [7, p. 4].

4 Implementation Details

We developed a set of routines that compute the following quantities using the MAGMA library (release 1.2.1):

- Variance–covariance matrix C
- κ_{LS}, condition number of x
- κ_{CW}, vector of the κ_i, condition numbers of the solution components

- $\bar{\kappa}_{LS}$, statistical estimate of κ_{LS}
- $\bar{\kappa}_{CW}$, vector of the statistical estimates κ_i

The variance–covariance computation requires inverting a triangular matrix and multiplying this triangular matrix by its transpose (similarly to the LAPACK routine DPOTRI [1, p. 26] that computes the inverse of a matrix from its Cholesky factorization). These operations use a block algorithm, which, for the diagonal blocks, is performed recursively. The recursive part is performed by the CPU for sake of performance while the rest of the algorithm is executed on the GPU.

The computation of the exact condition number κ_{LS} from the variance–covariance using Eq. (4) involves the computation of the spectral norm of C which is generally computed via an SVD. However, since A is a full-rank matrix, C is symmetric positive definite and its singular values coincide with its eigenvalues. Then we use an eigenvalue decomposition of C which is faster than an SVD because it takes into account the symmetry of C. The tridiagonalization phase is performed on the GPU while the subsequent eigenvalue computation is performed on the CPU host.

The statistical estimates require the generation and orthonormalization of random vectors followed by two triangular solves. The random generation and the triangular solves are performed on the GPU. The orthonormalization is performed on the CPU because it is applied to small matrices (small number of samples).

5 Performance Results

Our experiments have been achieved on a multicore processor Intel Xeon E5645 (2 sockets × 6 cores) running at 2.4 GHz (the cache size per core is 12 MB and the size of the main memory is 48 GB). This system hosts two GPU NVIDIA Tesla C2075 running at 1.15 GHz with 6 GB memory each. MAGMA was linked with the libraries MKL 10.3.8 and CUDA 4.1, respectively, for multicore and GPU. We consider random LLS problems obtained using the method given in [9] for generating LLS test problems with known solution x and residual norm.

We plot in Fig. 1, the CPU time to compute LLS solution and condition numbers using 12 threads and 1 GPU. We observe that the computation of the variance–covariance matrix and of the components conditioning κ_i are significantly faster than the cost for solving the problem with respectively a time factor larger than 3 and 2, this factor increasing with the problem size. The κ_i are computed using the variance-covariance matrix via Eq. (5). The time overhead between the computation of the κ_i and the variance–covariance computation comes from the computation of the norms of the columns (routine cublasDnrm2) which has a nonoptimal implementation.

As expected, the routines SCE_LLS and SCE_LLS_CW that compute statistical condition estimates for the solution and all solution components, respectively, outperform the other routines. Note that we did not mention on this graph the performance for computing κ_{LS} using Eq. (4). Indeed this involves an eigenvalue decomposition of the variance–covariance matrix (MAGMA routine magma_dsyevd_gpu), which turns out to be much slower than the LLS solution (MAGMA routine

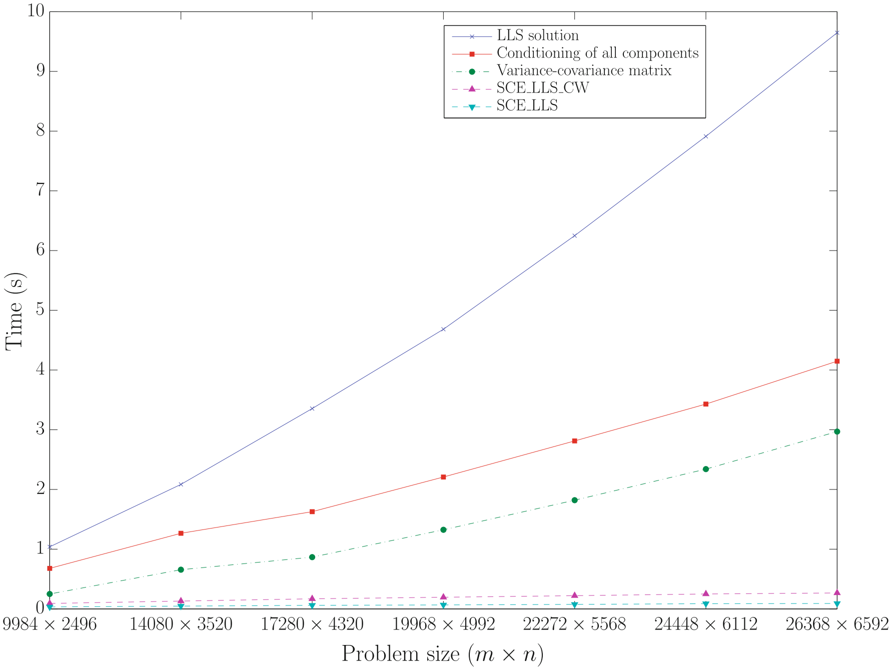


Fig. 1 Performance for computing LLS condition numbers with MAGMA

magma_dgels3_gpu) in spite of a smaller number of flops ($\mathcal{O}(n^3)$ vs $\mathcal{O}(mn^2)$) which shows that having an efficient implementation on the targeted architecture is essential to take advantage of the gain in flops.

We can illustrate this by comparing in Fig. 2 the time for computing an LLS solution and its conditioning using LAPACK and MAGMA. We observe that MAGMA provides faster solution and condition number but, contrary to LAPACK, the computation of the condition number is slower than the time for the solution, in spite of a smaller flop count. This shows the need for improving the Gflop/s performance of eigensolvers or SVD solvers for GPUs, but it also confirms the interest of considering statistical estimates on multicore-GPU architectures to get fast computations.

6 Conclusion

We proposed new implementations for computing LLS condition numbers using the software libraries LAPACK and MAGMA. The performance results that we obtained on a current multicore-GPU system confirmed the interest of using statistical

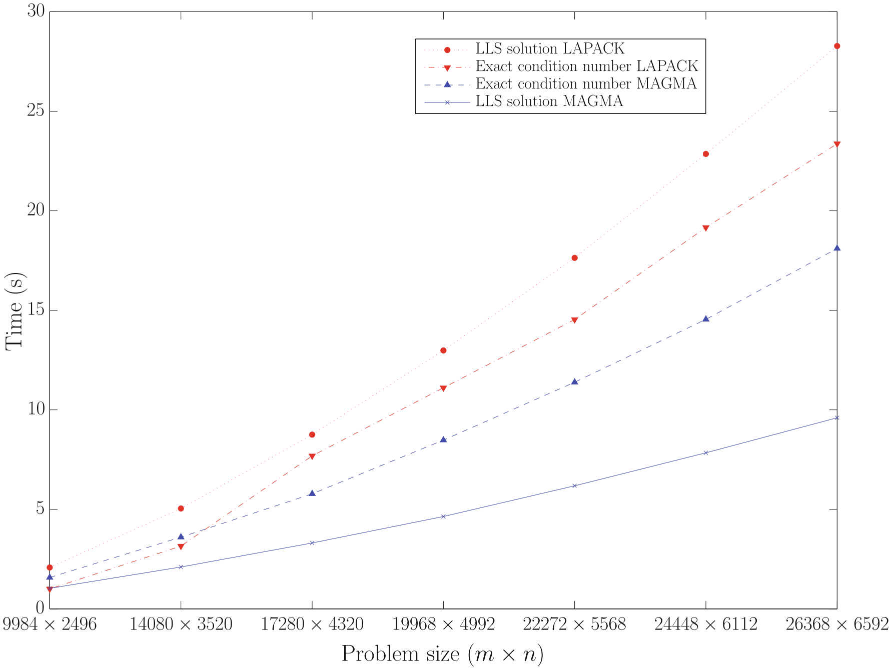


Fig. 2 Time for LLS solution and condition number

condition estimates. New routines will be integrated in the next releases of LAPACK and MAGMA to compute the variance–covariance matrix after a linear regression.

References

1. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J.D., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide, 3rd edn. SIAM, Philadelphia (1999)
2. Arioli, M., Baboulin, M., Gratton, S.: A partial condition number for linear least-squares problems. SIAM J. Matrix Anal. Appl. **29**(2), 413–433 (2007)
3. Baboulin, M., Dongarra, J., Gratton, S., Langou, J.: Computing the conditioning of the components of a linear least squares solution. Numer. Linear Algebra Appl. **16**(7), 517–533 (2009)
4. Baboulin, M., Dongarra, J., Tomov, S.: Some issues in dense linear algebra for multicore and special purpose architectures. In: 9th International Workshop on State-of-the-Art in Scientific and Parallel Computing (PARA'08), *Lecture Notes in Computer Science*, vol. 6126–6127. Springer-Verlag (2008)
5. Baboulin, M., Giraud, L., Gratton, S., Langou, J.: Parallel tools for solving incremental dense least squares problems. Application to space geodesy. J Algorithms Comput. Technol. **3**(1), 117–133 (2009)

6. Baboulin, M., Gratton, S., Lacroix, R., Laub, A.J.: Statistical estimates for the conditioning of linear least squares problems. In: Proceedings of 10th International Conference on Parallel Processing and Applied Mathematics (PPAM 2013) (2013)
7. Björck, A.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
8. Kenney, C.S., Laub, A.J., Reese, M.S.: Statistical condition estimation for linear least squares. SIAM J. Matrix Anal. Appl. **19**(4), 906–923 (1998)
9. Paige, C.C., Saunders, M.A.: LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Softw. **8**(1), 43–71 (1982)
10. Tomov, S., Dongarra, J., Baboulin, M.: Towards dense linear algebra for hybrid GPU accelerated manycore systems. Parallel Comput. **36**(5&6), 232–240 (2010)
11. Zelen, M.: Linear estimation and related topics. In: Todd, J. (ed.) Survey of Numerical Analysis, pp. 558–584. McGraw-Hill, New York (1962)

Coupled Spin Torque Nano-Oscillators: Stability of Synchronization

K. Beauvais, A. Palacios, R. Shaffer, J. Turtle, V. In and P. Longhini

Abstract In this work we explore the use of spin torque nano-oscillators (STNOs) to produce a spintronics voltage oscillator in the microwave range. STNOs are quite small—on the order of 100 nm—and frequency agile. However, experimental results till date have produced power outputs that are too small for practical use. We attempt to increase power output by investigating the dynamics of a system of electrically-coupled STNOs. Transverse Lyapunov exponents are used to quantitatively measure the local stability of synchronized limit cycles. The synchronized solution is found to be stable for a large region of two-parameter space. However, a two-parameter bifurcation diagram reveals a competing out-of-phase solution, causing bistability.

1 Introduction

Spin torque nano-oscillators (STNO) are a ferromagnet-based electronics component. In certain steady states, the magnetic moment precesses causing component resistance to oscillate [12]. Based on this oscillating resistance, an STNO can be utilized as a microwave-range voltage oscillator (see Fig. 1). However, STNOs tested till date have yet to produce adequate power. STNOs need to output at least 1 mW to be applicable [11]. The microwave power generated by an STNO was first measured in 2010 on the order of 1 nW [5]. One solution to increasing power is to electrically couple multiple oscillators. However, in experiments it has been proven that it is difficult to synchronize even two STNOs [8]. Thus, we have begun to study the dynamics of coupled STNOs to determine conditions for synchronization. In this chapter we describe the model in Cartesian coordinates and then project to complex

A. Palacios (✉) · K. Beauvais · R. Shaffer · J. Turtle
Nonlinear Dynamical Systems Group, Department of Mathematics,
San Diego State University, San Diego, CA 92182, USA
e-mail: apalacios@mail.sdsu.edu

V. In · P. Longhini
Space and Naval Warfare Systems Center, Code 2363, 53560 Hull Street,
San Diego, CA 92152-5001, USA

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_7

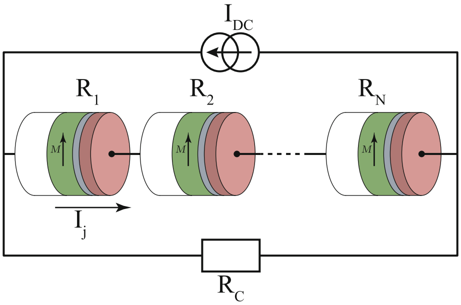


Fig. 1 Series arrayed STNOs with input current I_{DC} and output resistance R_c. The fixed ferromagnetic layer is *green* with magnetic moment **M**. The free ferromagnetic layer is *red* and its magnetic moment is modeled by **m**

stereographic coordinates. Next, we numerically analyze the local stability of synchronized limit cycles using transverse Lyapunov exponents (TLE). Following TLEs, we investigate global behavior by creating a two-parameter bifurcation diagram.

2 The Model

Magnetization in the free ferromagnetic layer is described by the Landau–Lifshitz equation with Gilbert damping and Slonczewski–Berger spin-torque term (LLGS) [1, 2, 4, 6, 13]

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \overbrace{-\gamma \mathbf{m} \times \mathbf{H}_{\text{eff}}}^{\text{precession}} + \overbrace{\lambda \mathbf{m} \times \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}}^{\text{damping}} - \overbrace{\gamma a\, g(P, \mathbf{m} \cdot \mathbf{M})\, \mathbf{m} \times (\mathbf{m} \times \mathbf{M})}^{\text{spin transfer torque}}, \tag{1}$$

where **m** represents the magnetization of the free ferromagnetic layer in Cartesian coordinates, γ is the gyromagnetic ratio, and $\mathbf{H}_{\text{eff}}$ is the effective external field. λ serves as the magnitude of the damping term. In the spin-torque term, a has units Oe and is proportional to the electrical current density [10]. g is a scalar function of the polarization factor P, **m**, and the fixed-layer magnetization direction **M**. To determine the change of field direction with respect to time, we must consider three different classes of torques acting on the field direction **m**: effective external magnetic field $\mathbf{H}_{\text{eff}}$, damping λ, and spin transfer torque. $\mathbf{H}_{\text{eff}}$ is the sum of several factors that can be effectively represented as external fields. The factors that we consider in this fashion are exchange, anisotropy, and demagnetization. The actual external, or applied, field rounds out the sum

$$\mathbf{H}_{\text{eff}} = \mathbf{H}_{\text{exchange}} + \mathbf{H}_{\text{anisotropy}} + \mathbf{H}_{\text{demagnetization}} + \mathbf{H}_{\text{applied}}.$$

We model the free layer as a single particle who's magnetization **m** represents the average of the layer. Thus, there is no exchange with adjacent magnetic moments $\mathbf{H}_{\text{exchange}} = 0$.

Complex Stereographic Projection

A spherical surface can be projected onto a plane by using the complex variable ω and the following relationships:

$$\omega = \frac{m_x + i m_y}{1 + m_z} \quad \Rightarrow \quad \mathbf{m} = \begin{bmatrix} \dfrac{\omega + \bar{\omega}}{1 + |\omega|^2} \\ -i\dfrac{(\omega - \bar{\omega})}{1 + |\omega|^2} \\ \dfrac{1 - |\omega|^2}{1 + |\omega|^2} \end{bmatrix}. \tag{2}$$

Here $\bar{\omega}$ is the complex conjugate of ω. This projection maps the sphere's north pole to the origin and the south pole to infinity. Building on [9, 10], we reduce Eq. (1) to the form

$$\begin{aligned} \dot{\omega} =& \frac{\gamma}{1 - i\lambda}\Big(- a\omega + i h_{a3}\omega + \frac{h_{a2}}{2}(1 + \omega^2) \\ & + i m_{\|}\kappa \left[\cos\theta_{\|}\omega - \frac{1}{2}\sin\theta_{\|}\left(\mathrm{e}^{i\phi_{\|}} - \omega^2 \mathrm{e}^{-i\phi_{\|}}\right)\right] \\ & - \frac{i4\pi S_o}{(1 + |\omega|^2)}\Big[N_3(1 - |\omega|^2)\omega - \frac{N_1}{2}(1 - \omega^2 - |\omega|^2)\omega \\ & \qquad - \frac{N_2}{2}(1 + \omega^2 - |\omega|^2)\omega - \frac{(N_1 - N_2)}{2}\bar{\omega}\Big]\Big), \end{aligned} \tag{3}$$

where h_{a2} is the magnitude of the applied field in the y-direction and h_{a3} is the magnitudes of the applied field in the z-direction. κ is the anisotropy magnitude who's direction is determined by the spherical-coordinate parameters $\theta_{\|}$ and $\phi_{\|}$. The anisotropy is scaled by $m_{\|} = \mathbf{m} \cdot \mathbf{e}_{\|}$ where

$$\mathbf{e}_{\|} = \begin{bmatrix} \sin\theta_{\|}\cos\phi_{\|} \\ \sin\theta_{\|}\sin\phi_{\|} \\ \cos\theta_{\|} \end{bmatrix}.$$

S_0 is the saturation magnetization. Finally, N_1, N_2, and N_3 describe the effective demagnetization field resulting from the shape of the free layer and are constrained by the relationship $N_1 + N_2 + N_3 = 1$. The magnetic moment of an STNO is now described by two dimensions and in a polynomial-like form.

Coupling

Coupling is achieved by modeling a simple electrical circuit with STNOs arrayed in series or parallel. Figure 1 depicts the series configuration. The resistance of each STNO R_i is a function of the angle θ_i between M (fixed layer-green) and $\mathbf{m}$ (free layer-red):

$$R_i = R_{0i} - \Delta R_i \cos\theta_i.$$

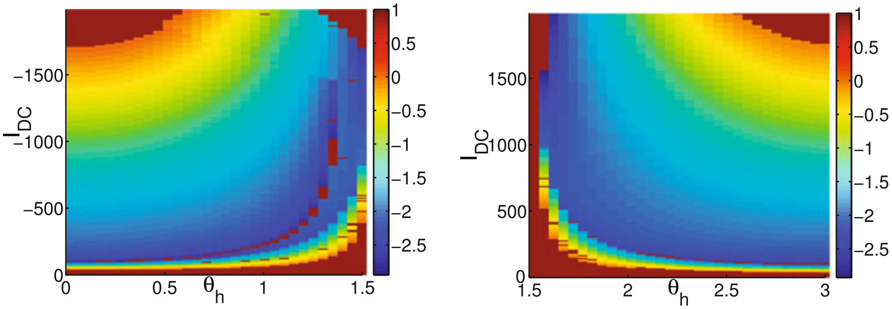


Fig. 2 Sum of the transverse Lyapunov exponents calculated over two parameters, the electrical current I_{DC} and the angle of the applied field θ_h, with the applied field varying from 0 to $\frac{\pi}{2}$ (*left*) and from $\frac{\pi}{2}$ to π (*right*). Extreme *red* indicates no oscillations

Here, R_0 is the median resistance of an STNO and ΔR is the maximum variance in resistance.

3 Transverse Lyapunov Exponents

Transverse Lyapunov exponents allow us to quantify the local stability of a synchronized orbit [3, 7]. Specifically, TLEs use the linearized system to measure how a small perturbation transverse to the synchronization manifold ($z_s = z_1 = \cdots = z_n$) grows or contracts. Figure 2 depicts the result of numerically calculating TLEs for two serially-coupled and identical STNOs. Here we vary the electrical current I_{DC} with a grid step size $\Delta I_{DC} = 10$ and the applied field angle θ_h with a step size $\Delta\theta_h = 0.05$. Initial conditions are on the synchronization manifold and close to the expected steady state. In this case, the sum of TLEs is a good representation of stability, hence the plot is color coded accordingly. A negative sum indicates stable while a positive sum indicates instability of the synchronized orbit. Additionally, the uniform red color represents areas where no oscillations were detected. These results indicate large regions of parameter space where the synchronized solution is stable. Further, there are distinct boundaries of oscillations that may trace the locus of Hopf bifurcations.

4 Numeric Bifurcation Diagram

We have shown that synchronized oscillations are locally stable for a large parameter space. However, simulations in the same parameter space using random initial conditions show out-of-phase steady-state behaviors [14]. Using the software package XPPAUT, we have created a two-parameter bifurcation diagram (Fig. 3) that

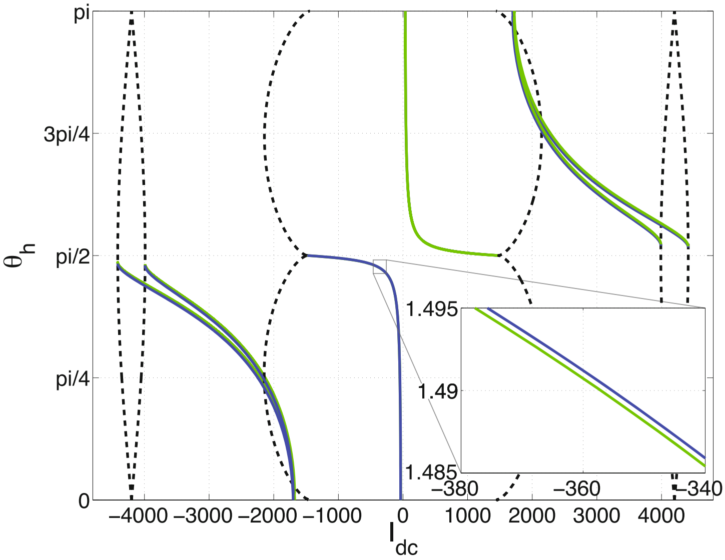


Fig. 3 Two-parameter bifurcation diagram plotting input current I_{DC} v applied field angle θ_h. *Dashed black lines* trace saddle-node bifurcations, while *solid lines* are Hopfs. *Green solid lines* are Hopfs that spawn synchronized limit cycles and *solid blue lines* are Hopfs that spawn out-of-phase oscillations

includes the parameter spaces from Fig. 2. The diagram reveals a number of back-to-back Hopf bifurcations that are consistent with the boundary of oscillations in Fig. 2. Additionally, each pair of back-to-back Hopfs spawns one synchronized and one out-of-phase limit cycle. One-parameter bifurcation diagrams in I_{DC} (θ_h fixed) indicate that the out-of-phase solution is also locally stable for most of the oscillating region. Hence, the region of interest exhibits bistability where we expect the initial conditions to determine synchronized or out-of-phase steady-state behavior.

5 Remarks

The LLGS Eq. (1) is a nonlinear first-order ordinary differential equation confined to the unit sphere $\|\mathbf{m}\|_2 = 1$. We are able to reduce the dimension of the system one third using complex-stereographic coordinates. Not only does this increase the efficiency of numerics but also simplifies integration by fixing the magnitude of $\mathbf{m}$ by the nature of the coordinate system.

The calculation of TLEs in Fig. 2 gives the positive result of stable synchronized oscillations in a large parameter space. However, the TLE measurement is inherently local, and therefore does not necessarily reflect global behavior. Using XPPAUT to create a two-parameter bifurcation diagram, we discover the existence of an out-of-phase limit cycle that is also locally stable. This bistability indicates that synchronization can only be achieved if the initial conditions fall within the basin of attraction of the synchronized solution.

In future work we are interested in calculating the basins of attraction, but we are also interested in the behavior of the system for many more STNOs. As we increase the number of oscillators N, we expect a nonlinear increase in the number of oscillatory steady states. We intend to leverage the symmetry-group representation of the coupling to predict the type, existence, and stability of out-of-phase oscillations. Combined with the TLE computation, this will allow us to determine if a region exists where the synchronized solution is globally stable.

References

1. Berger, L.: Emission of spin waves by a magnetic multilayer traversed by a current. Phys. Rev. B **54**(13), 9353–9358 (1996). doi:10.1103/PhysRevB.54.9353
2. Bertotti, G., Mayergoyz, I., Serpico, C.: Analytical solutions of Landau-Lifshitz equation for precessional dynamics. Phys. B **343**(1–4), 325–330 (2004)
3. Chitra, R., Kuriakose, V.: Phase effects on synchronization by dynamical relaying in delay-coupled systems. Chaos: Interdiscip. J. Nonlinear Sci. **18**(2), 023,129–023, 129 (2008)
4. d'Aquino, M.: Nonlinear magnetization dynamics in thin-films and nanoparticles. Ph.D. thesis, Universitá degli Studi di Napoli Federico II, Naples, Italy (2004)
5. Demidov, V., Urazhdin, S., Demokritov, S.: Direct observation and mapping of spin waves emitted by spin-torque nano-oscillators. Nat. Mater. **9**(12), 984–988 (2010)
6. Gilbert, T.: A phenomenological theory of damping in ferromagnetic materials. IEEE Trans. Magn. **40**(6), 3443–3449 (2004)
7. Krasovskiĭ, N.N.: Stability of Motion: Applications of Lyapunov's Second Method to Differential Systems and Equations with Delay. Stanford University Press (1963)
8. Li, D., Zhou, Y., Zhou, C., Hu, B.: Global attractors and the difficulty of synchronizing serial spin-torque oscillators. Phys. Rev. B **82**(14), 140,407 (2010)
9. Murugesh, S., Lakshmanan, M.: Bifurcation and chaos in spin-valve pillars in a periodic applied magnetic field. Chaos **19**, 043,111 (2009)
10. Murugesh, S., Lakshmanan, M.: Spin-transfer torque induced reversal in magnetic domains. Chaos Solitons Fractals **41**, 2773–2781 (2009)
11. Persson, J., Zhou, Y., Akerman, J.: Phase-locked spin torque oscillators: impact of device variability and time delay. J. Appl. Phys. **101**(9), 09A503 (2007)
12. Slavin, A., Tiberkevich, V.: Nonlinear auto-oscillator theory of microwave generation by spin-polarized current. IEEE Trans. Magn. **45**(4), 1875–1918 (2009)
13. Slonczewski, J.C.: Current-driven excitation of magnetic multilayers. J. Magn. Magn. Mater. **159**(1–2), L1–L7 (1996)
14. Turtle, J;.A.: Numerical exploration of the dynamics of coupled spin torque nano oscillators. M.S. thesis, San Diego State University, San Diego, CA (2012).

Nonlinear Robust Control and Regulation Problems for Biomedical Dynamical Systems

Aziz Belmiloudi

Abstract Motivated by topics and issues critical to human health, and safety and efficacy of medical treatment practices, this communication investigates a nonlinear robust control approach of some uncertain biomedical nonlinear complex systems. The concept consists in setting the problem in the worst-case disturbances, which leads to the game theory in which controls and disturbances (which destabilize the dynamical behavior of the system) play antagonistic roles. The proposed strategy consists in controlling these instabilities by acting on certain parameters and data to maintain the system in a desired state (see *Stabilization, Optimal and Robust Control*, Springer, London (2008)). This approach is applied to two problems: first, controlling and regulating the blood glucose level in subjects with type 1 diabetes and predicting the dosages of insulin administered, and second, controlling and stabilizing the thermal distribution and damage during the treatment of cancer, in order to eradicate tumor while preserving the surrounding health tissues.

1 Introduction and Outline

The problem studied in this chapter derives from the modeling, stabilizing control, and regulation of the dosage of drug and thermal required for optimal therapy of various diseases and injuries of the human body. To ensure effective treatment and improve the lives of patients, it is necessary not only to have reliable mathematical models, but also, despite the complexity of the systems, to have nonlinear control methods capable to ensure safety and stability under all circumstances with a robust stability and performance. Consequently, this has greatly emphasized the need for sophisticated mathematical models of dynamic systems and methodologies capable of predicting, understanding, and optimizing different complex phenomena occurring in these fields, despite different sources of uncertainty like the absence of complete or reliable data, neglected dynamics, or intrinsic physical variability. The challenge here is, e.g., to reduce the uncertainty and increase the reliability of model

A. Belmiloudi (✉)
IRMAR-INSA of Rennes, 20 avenue Buttes de Coësmes, 35700 Rennes, France
e-mail: aziz.belmiloudi@math.cnrs.fr; aziz.belmiloudi@insa-rennes.fr

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_8

predictions about the performance of these complex and realistic systems. Motivated by the above discussion, the goal of our contribution is to study time-dependent identification, regulation, and stabilization problems related to the nonlinear phenomena of transport and transformation occur in biomedical domain, by using new and modern robust control theory developed in [2]. For this, we consider two important applications, which refer to two high impact diseases namely cancer with thermotherapy and diabetes with insulin therapy. Thermal conductivities in living tissues and glucose metabolism are nonlinear and very complex processes which use different phenomenological mechanisms including conduction, convection, metabolism, evaporation, etc., and subject to various perturbations and physiological and pathophysiological variations. An analysis of automated-treatment taking into account these parameters will be very beneficial for dosage distribution, treatment planning and control of the treatment outcome. The organization of the paper is as follows. In Sect. 2, we present our approach of the modern robust control theory. In Sect. 3 we study the two diseases: the controlling and regulating of the blood glucose level in subjects with type 1 diabetes and predicting the dosages of insulin administered, and second, controlling and stabilizing the thermal dose distribution in tissue and damage during thermotherapy.

2 From Models to Regulation Problems: Terminology and Process

To predict the response of dynamic systems from given parameters, data and source terms requires a mathematical model of the behavior of the process under investigation and a physical theory linking the state variables of the model to data and parameters. This prediction of the observation constitutes the so-called direct (or primal) problem and it is usually defined by one or more coupled integral, ordinary or partial differential systems and sufficient boundary and initial conditions for each of the main fields (such as temperature, concentration, velocity, wave, etc.). Initial and/or boundary conditions are essential for the design and characterization of any model systems. If any of the conditions necessary to define a direct problem are unknown or rather badly known, an *inverse* (*or control*) *problem* results. The resolution of the inverse problems thus provides them essential information, which is necessary to the comprehension of the various processes which can intervene in these models. This resolution needs some partial information of some unknown parameters and fields (observations) given, for example, by experiment measurements. The inverse problem is used for systems where uncertainties are neglected. But it is well known that many uncertainties occur in the most realistic studies, e.g., of life science problems. The presence of these uncertainties may induce complex behaviors, e.g., oscillations, instability, bad performances, etc. So, if uncertainties, stability, and performance validation occur, a *robust control problem* results. The fundament of robust control theory is to take into account these uncertain behaviors and to analyze how the control system can deal with this problem. The uncertainty can be of two types: first, the errors (or imperfections) coming from the model and, second, the unmeasured noises and fluctuations that act on the systems. The goal of robust control theory is to control these instabilities, either by acting on some parameters to maintain the system in a desired state (target), or by calculating the limit of these parameters before the system becomes unstable ("predict to act"). Our robust control approach consists in setting the problem in the worst-case disturbances which leads to the *game theory* in which controls and disturbances play antagonistic roles. For more details on this new approach and its application to different models describing realistic physical and biological process, see the book [2]. The essential data used in our robust control approach are the following.

- A known nonlinear operator $\mathcal{G}$ which represents the dynamical system:

$$\partial_t U + \mathcal{G}(x,t; f, g, U) = 0, \ \ \textit{with the initial condition} : U(t=0) = U_0, \tag{1}$$

where (x,t) are the space-time variables, $\partial_t U$ denotes $\frac{\partial U}{\partial t}$, $(U_0, f, g) \in \mathcal{X}$ is the input of the system ((f,g) is, e.g., boundary conditions, source terms, parameters and others), and $U \in \mathcal{Z}$ is the state of the system, where $\mathcal{X}$ and $\mathcal{Z}$ are two spaces of input data and output solutions, respectively.

- A "control" variable φ in $U_{ad} \subset \mathcal{U}_1$ (set of "admissible controls") and a "disturbance" variable $\psi = (\psi_1, \psi_2)$ in $V_{ad} \subset \mathcal{U}_2 = \mathcal{U}_{21} \times \mathcal{U}_{22}$ (set of "admissible disturbances"), with $\mathcal{U}_1$ and $\mathcal{U}_2$ two spaces of controls and disturbances, respectively.
- For a chosen control-disturbance (φ, ψ), the perturbation problem, which models fluctuations $(\psi_1, \varphi, \psi_2, u)$ to the target (U_0, f, g, U), is given by

$$\partial_t u + \tilde{\mathcal{G}}(x,t; \varphi, \psi_2, u) = 0, \ \ \textit{with} \ \ u(t=0) = \psi_1, \tag{2}$$

where the perturbation of the model $\mathcal{G}$ is given by $\tilde{\mathcal{G}}(.; \varphi, \psi_2, u) = \mathcal{G}(.; f + \mathcal{B}_1\varphi, g + \mathcal{B}_2\psi_2, U + u) - \mathcal{G}(.; f, g, U)$, and $\mathcal{B}_1$ (respectively $\mathcal{B}_2$) is a bounded linear operator from $\mathcal{U}_1$ (respectively $\mathcal{U}_2$) into $\mathcal{Z}$. In the sequel we denote by $u = \mathcal{F}(\varphi, \psi)$ the solution of the direct problem (2).

- An "observation" u_{obs} which is supposed to be known (for example the desired tolerance for the perturbation or the offset given by measurements).
- A "cost" (or "objective") functional $\mathcal{J}$ which is defined from a real-valued and positive function $\mathcal{L}$ by $\mathcal{J}(\varphi, \psi) := \mathcal{L}((\varphi, \psi); \mathcal{F}(\varphi, \psi))$. The goal is to find a saddle point of $\mathcal{J}$, i.e., a solution $(\varphi^*, \psi^*) \in U_{ad} \times V_{ad}$ (subject to (2)) of

$$\mathcal{J}(\varphi^*, \psi) \le \mathcal{J}(\varphi^*, \psi^*) \le \mathcal{J}(\varphi, \psi^*), \ \ \forall(\varphi, \psi) \in U_{ad} \times V_{ad}. \tag{3}$$

- Then, we have to determine the gradient of $\mathcal{J}$ and the necessary conditions of optimality by differentiating $\mathcal{F}$ and introducing an *adjoint model*. The adjoint problem is in the form ($\mathcal{G}^*$ is linear on the state $\tilde{u}$)

$$-\partial_t \tilde{u} + \mathcal{G}^*(\varphi, \psi, \mathcal{F}(\varphi, \psi), u_{obs}; \tilde{u}) = 0, \ \ \textit{with final condition} : \tilde{u}(t=T) = \tilde{u}_T. \tag{4}$$

- Define an algorithm allowing to solve numerically the control problem.

3 Biomedical Applications: Two High Impact Diseases

In clinical practice, measurements, material data, behavior of patients, and other process are highly disturbed and affected by noises and errors. Consequently, in order to obtain a solution robust to the noises and fluctuations, it is necessary to incorporate theses in the modeling and to analyze the robust regulation of the deviation of the model from the desired dose distribution target, due to fluctuations. In this section, we formulate the robust control problem in the case of two high impact diseases namely diabetes and cancer therapies.

3.1 Blood Glucose and Type 1 Diabetic Patient

The goal here is to regulate and stabilize the injection of insulin via blood glucose sensor. This is motivated by the development of reliable and feasible control strategies for patients which automatically connect continuous glucose sensor and insulin injection, without patient intervention. For this we consider Bergman type model

$$d_t X = \mathcal{M}(t, X) + F(t) \;\; in \; [0, T], \;\; with \;\; X(0) = X_0, \tag{5}$$

where $X = (G, H, I, U)$, d_t denotes $\frac{d}{dt}$, G is the blood glucose, H is the remote insulin, I is the blood insulin, U is the insulin in the skin and $F(t) = (h(t), 0, 0, f(t))$ with h, e.g., the glucose flow due to the consumption of a meal and f the injected insulin flow. The operator $\mathcal{M} = (\mathcal{M}_i)_{i=1,4}$ is such that

$$\begin{aligned} &\mathcal{M}_1(t, X) = -P_1(G - G_b) - H\,G + h,\; \mathcal{M}_2(t, X) = -P_2 H + P_3(I - I_b),\\ &\mathcal{M}_3(t, X) = -P_4\, I + P_5\, U,\; \mathcal{M}_4(t, X) = -P_6\, U + f, \end{aligned} \tag{6}$$

where $P_i, i = 1, 6$ are system parameters and G_b (resp. I_b) is a base value of plasma glucose (respectively insulin). All data are assumed to be in $L^\infty(0, T)$. For some details about mathematical modeling, see e.g., [5, 6]. The well-posedness of system (5) can be obtained by the contraction of the operator $X \longrightarrow X_0 + \int_0^T (\mathcal{M}(t, X) + F(t))dt$, the Lipschitz condition and linear growth of $\mathcal{M}$ and Gronwall lemma. Now we formulate the robust control problem. First, we develop the perturbation problem, which models fluctuations x to the target X, i.e., we assume that X satisfies (5) and (6) with data (X_0, f, h) and $X + x$ satisfies (5) and (6) with the data $(X_0 + x_0, f + \varphi, h + \psi)$. Hence, we consider the following system in $[0, T]$ (where $K = P_1 + H$)

$$\begin{aligned} &d_t x_1 = -K\, x_1 - G\, x_2 - x_1\, x_2 + \psi,\;\; d_t x_2 = -P_2\, x_2 + P_3\, x_3,\\ &d_t x_3 = -P_4\, x_3 + P_5\, x_4,\;\; d_t x_4 = -P_6 x_4 + \varphi,\\ &x(0) = x_0,\\ &under\ pointwise\ constraints: \tau_1 \le \psi \le \tau_2,\;\; and\;\; \delta_1 \le \varphi \le \delta_2. \end{aligned} \tag{7}$$

Let now $\mathcal{V}_{ad} = \{\psi \in L^2(0,T) : \tau_1 \leq \psi \leq \tau_2\}, \mathcal{U}_{ad} = \{\varphi \in L^2(0,T) : \delta_1 \leq \varphi \leq \delta_2\}$, the operator solution of (7) denote by $\mathcal{F} : (\varphi,\psi) \in \mathcal{U}_{ad} \times \mathcal{V}_{ad} \longrightarrow \ x = \mathcal{F}(\varphi,\psi) \in \mathcal{Z}$ and observation G_{obs} be given (difference between a measurement reading and the true value of that measurement). Our problem is to find a saddle point (φ^*, Ψ^*) of $\mathcal{J}(\varphi,\psi) = \frac{1}{2} \parallel x_1 - G_{obs} \parallel^2_{L^2(0,T)} + \frac{\alpha}{2} \parallel \varphi \parallel^2_{L^2(0,T)} - \frac{\beta}{2} \parallel \psi \parallel^2_{L^2(0,T)}$ subject to (7), with $\alpha > 0$ and $\beta > 0$. The arguments of [2] extend directly to the present work. So, we have the following existence result and first-order optimality conditions.

Theorem 1 *For α and β sufficiently large, there exists an optimal solution $(\varphi^*, \Psi^*) \in \mathcal{U}_{ad} \times \mathcal{V}_{ad}$ and $x^* = \mathcal{F}(\varphi^*, \Psi^*) \in \mathcal{Z}$ such that (φ^*, Ψ^*) is a saddle point of $\mathcal{J}$. Moreover (φ^*, ψ^*, x^*) can be characterized by*

$$\begin{aligned}
&\frac{\partial \mathcal{J}}{\partial \varphi}(\varphi^*, \Psi^*).(\varphi - \varphi^*) = \int_0^T (\tilde{x}_4 + \alpha\varphi^*)(\varphi - \varphi^*)dt \geq 0, \ \forall \varphi \in \mathcal{U}_{ad} \\
&\frac{\partial \mathcal{J}}{\partial \Psi}(\varphi^*, \Psi^*).(\Psi - \Psi^*) = \int_0^T (\tilde{x}_1 - \beta\psi^*)(\psi - \psi^*)dt \leq 0, \ \forall \psi \in \mathcal{V}_{ad}
\end{aligned} \tag{8}$$

where $\tilde{x} = \tilde{\mathcal{F}}(\varphi^*, \Psi^*)$ is the solution of the adjoint model

$$\begin{aligned}
&-d_t\tilde{x}_1 = -(K + x_2^*)\tilde{x}_1 + (x_1^* - G_{obs}), \ \ -d_t\tilde{x}_2 = -(G + x_1^*)\tilde{x}_1 - P_2\tilde{x}_2, \\
&-d_t\tilde{x}_3 = P_3\tilde{x}_2 - P_4\tilde{x}_3, \ \ -d_t\tilde{x}_4 = P_5\tilde{x}_3 - P_6\tilde{x}_4, \ \ with \ \ \tilde{x}(T) = 0.
\end{aligned} \tag{9}$$

3.2 Temperature Distribution and Cancer Therapy

The goal here is to regulate the effects of thermal physical properties on the transient temperature of tissues via the online temperature measurements by magnetic resonance imaging. This is motivate by the fact that heating the cell up to high temperatures, the tumor cells don't repair themselves as well, hence they are more susceptible to the effects of thermotherapy. To treat the system of motion in living body, we consider the transient bioheat transfer type model in a form as follows

$$c(x)\partial_t U = div(\kappa(x)\nabla U) - P(U - U_a) - (\vec{\vartheta}.\nabla)U + f + g \ in \ Q = \Omega \times (0,T),$$

subjected to the heat-flux boundary condition and the initial condition

$$\begin{aligned}
&(\kappa\nabla U).\mathbf{n} \ = -q(U - U_b) - \lambda(x)(L(U) - L(U_b)) + h \ \ in \ \Sigma = \partial\Omega \times (0,T), \\
&U(0) = U_0 \ \ in \ \ \Omega,
\end{aligned} \tag{10}$$

where U is the temperature distribution, $L(U) = \mid U \mid^3 U$, $T > 0$ is a given final time, the body Ω is an open bounded domain in $\mathbb{R}^m$, $m \leq 3$ with a smooth boundary $\Gamma = \partial\Omega$, $\mathbf{n}$ is the unit outward normal to Γ, $P \in L^\infty(Q)$ is the blood perfusion rate and $q \in L^\infty(\Sigma)$ is the heat transfer coefficient. The conductivity of tissue κ satisfies $\nu \geq \kappa = \sigma^2 \geq \mu > 0$ (where ν and μ are constants). The term $\vec{\vartheta}$ is the blood flow velocity. The term f is a distributed energy source such as focused ultrasound and laser beams, and g is the energy generated by the metabolic processes. The term h is the heat flux due to evaporation. The functions $U_a \in L^\infty(\mathcal{Q})$ and $U_b \in L^\infty(\Sigma)$ are the blood and the bolus temperatures, respectively. The term $\lambda = \sigma_B \epsilon_e$ is assumed to be in $L^\infty(\Gamma)$ where σ_B is Stefan–Bolzmann's constant and ϵ_e is the effective emissivity. For more details about modeling and mathematical analysis see [1, 3, 4]. Now we formulate the robust control problem. First, we introduce the perturbation problem, which models fluctuations u to the target temperature U. Assume that U satisfies (10) with data (U_0, P, f, g, h) and $U + u$ satisfies (10) with data $(U_0 + u_0, P + \varphi_1, f + \varphi_2, g + \psi_1, h + \psi_2)$. Then, u satisfies (where $u_a = U - U_a$)

$$\begin{aligned}
&c(x)\partial_t u - div(\kappa(x)\nabla u) = -\varphi_1(u - u_a) - Pu - (\vec{\vartheta}.\nabla)u + \varphi_2 + \psi_1 \;\; in \;\; \mathcal{Q},\\
&(\kappa(x)\nabla u).\mathbf{n} \; = -qu - \lambda(x)(L(u + U) - L(U)) + \psi_2 \;\; in \;\; \Sigma,\\
&u(0) = u_0 \;\; in \;\; \Omega,\\
&under\ pointwise\ constraints\ \tau_1 \le \varphi_1 \le \tau_2 \;\; a.e.\ in\ \mathcal{Q}.
\end{aligned} \tag{11}$$

Let $(\mathcal{K}_i)_{i=1,2}$ (resp. $\mathcal{K}_3$) be convex, closed, nonempty and bounded subset of $L^2(\mathcal{Q})$ (resp. $L^2(\Sigma)$), $\mathcal{U}_{ad} = \{\varphi \in L^2(\mathcal{Q}) : \tau_1 \le \varphi \le \tau_2 \; in \; \mathcal{Q}\} \times \mathcal{K}_1$, $\mathcal{V}_{ad} = \mathcal{K}_2 \times \mathcal{K}_3$, $\mathcal{F} : (\phi, \Psi) \in \mathcal{U}_{ad} \times \mathcal{V}_{ad} \longrightarrow \; u = \mathcal{F}(\phi, \Psi) \in \mathcal{Z}$ be the operator solution of (11) and observation $\mathfrak{m}_{obs}$ be given (via MRI measurements). Our problem is to find a saddle point (X^*, Y^*) of $\mathcal{J}(\varphi, \psi) = \frac{1}{2} \parallel (\gamma u + \delta\varphi_1) - \mathfrak{m}_{obs} \parallel^2_{L^2(\mathcal{Q})} + \frac{\alpha}{2} \parallel \mathcal{N}\phi \parallel^2_{L^2(\mathcal{Q})} - \frac{\beta}{2} \parallel \mathcal{M}\Psi \parallel^2_{L^2(\mathcal{Q}) \times L^2(\Sigma)}$, subject to (11) (with $\alpha, \beta > 0$ constants and $\gamma, \delta > 0$ in $L^\infty(\overline{\mathcal{Q}})$), where the matrices $\mathcal{M} = diag(\sqrt{m_1}, \sqrt{m_2})$ and $\mathcal{N} = diag(\sqrt{n_1}, \sqrt{n_2})$ are such that $m_1 + m_2 \neq 0$, $n_1 + n_2 \neq 0$. The arguments of [2] extend directly to the present work. So, we have the following existence result and first-order optimality conditions.

Theorem 2 *For α and β sufficiently large, there exists an optimal solution $(\phi^*, \Psi^*) \in \mathcal{U}_{ad} \times \mathcal{V}_{ad}$ and $u^* = \mathcal{F}(\phi^*, \Psi^*) \in \mathcal{Z}$ such that (ϕ^*, Ψ^*) is a saddle point of $\mathcal{J}$. Moreover (ϕ^*, Ψ^*, u^*) can be characterized by (for all $(\phi, \Psi) \in \mathcal{U}_{ad} \times \mathcal{V}_{ad}$)*

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \phi}(\phi^*, \Psi^*).(\phi - \varphi^*) &= \int\!\!\int_{\mathcal{Q}} (\alpha n_1 \varphi_1^* + (u^* - u_a)\tilde{u} + \delta(M^* - \mathfrak{m}_{obs}))(\varphi_1 - \varphi_1^*) dxdt\\
&\quad + \int_{\Omega} (\alpha n_2 \varphi_2^* - \tilde{u})(\varphi_2 - \varphi_2^*) dxdt \ge 0,\\
\frac{\partial \mathcal{J}}{\partial \Psi}(\varphi^*, \Psi^*).(\Psi - \Psi^*) &= -\int\!\!\int_{\mathcal{Q}} (\tilde{u} + \beta m_1 \psi_1^*)(\psi_1 - \psi_1^*) \, dxdt\\
&\quad - \int\!\!\int_{\Sigma} (\tilde{u} + \beta m \psi_2^*)(\psi_2 - \psi_2^*) d\Gamma dt \le 0,
\end{aligned}$$

where $M^* = \gamma u^* + \delta \varphi_1^*$ and $\tilde{u} = \tilde{\mathcal{F}}(\phi^*, \Psi^*)$ is the solution of the adjoint problem

$$\begin{aligned}
&-c(x)\partial_t \tilde{u} - div(\kappa \nabla \tilde{u}) - div(\tilde{u}\vec{\vartheta}) + (\varphi_1^* + P)\tilde{u} + \gamma(M^* - \mathfrak{m}_{obs}) = 0 \;\; in \;\; \mathcal{Q},\\
&-\kappa \nabla \tilde{u}.\mathbf{n} \; = q\tilde{u} + \lambda(x) L'(u^* + U)\tilde{u} + \tilde{u}\vec{\vartheta}.\mathbf{n} \; in \, \Sigma,\\
&and\ the\ final\ condition\ \tilde{u}(T) = 0 \;\; in \;\; \Omega\\
&with\ L'(\bullet) = 4|\bullet|^3.
\end{aligned}$$

Remark 1 This chapter concerns the real-time control and robust stabilization problems for predicting and regulating the dosages of insulin and temperature administrate, via the online desired states provided by sensor measurements. It is clear that we can consider other control, disturbance and observation functions and obtain the same results by using the same techniques. For other models in realistic situations we can refer for thermal and insulin therapy, e.g., to [4] and [7] and references therein. For numerical resolution of robust control problems see [2].

References

1. Belmiloudi, A.: Analysis of the impact of nonlinear heat transfer laws on temperature distribution in irradiated biological tissues: mathematical models and optimal controls. J. Dynam. Control Syst. **108**, 217–254 (2007)
2. Belmiloudi, A.: Stabilization, Optimal & Robust Control: Theory & Applications in Biological & Physical Sciences. Springer-Verlag, London (2008)
3. Belmiloudi, A.: Parameter identification problems and analysis of the impact of porous media in biofluid heat transfer in biological tissues during thermal therapy. Nonlinear Anal. Real World Appl. **11**, 1345–1363 (2010)
4. Belmiloudi, A.: Thermal therapy: stabilization and identification. In: Belmiloudi, A. (ed.) Heat Transfer-Mathematical Modelling, Numerical Methods and Information Technology, pp. 33–76. INTECH, Vienna (2011)
5. Li, J., Johnson, D.: Mathematical models of subcutaneous injection of insulin analogues: a mini-review. Discrete Continuous. Dyn. Syst. B **12**, 401–414 (2009)
6. Makrogloua, A., Lib, J., Kuang, Y.: Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview. Appl. Numer. Math. **56**, 559–573 (2006)
7. Penel, M., Gueguen, H., Belmiloudi, A.: A robust receding horizon control approach to artificial glucose control for type 1 diabetes. In: 9th IFAC Symposium on Nonlinear Control Systems, pp. 833–838 (2013)

A Model of Heat and Water Transport in Frozen Porous Media and Fractured Rock Masses

Michal Beneš, Lukáš Krupička and Radek Štefan

Abstract In this contribution, the model of heat and water transport in frozen porous media and fractured rock masses in conditions of freezing and thawing is analyzed. We present results concerning the existence of the numerical solution. Numerical scheme is based on semi-implicit discretization in time. The spacial discretization is carried out by the finite element method (FEM) and it is implemented in MATLAB. We also present an illustrative numerical example.

1 Governing Equations

Phenomena involving partially frozen porous media or fractured rock masses are important in agriculture, civil or transport engineering, ecological and natural systems, and much attention is focused on the modeling of their behavior. According to the main physical processes in porous media under freezing–thawing conditions, some hypotheses are proposed, including that: (i) Darcy's law applies to water movement in both unfrozen and frozen soil, (ii) porous media is undeformable, (iii) the influence of soil water vapor migration on unfrozen water and heat flow transfers is ignored, (iv) all processes are single valued, i.e., hysteresis is not present in the characteristic curves, and (v) ice is immovable.

The governing equations of the model are as follows:

M. Beneš (✉) · L. Krupička
Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic
e-mail: benes@mat.fsv.cvut.cz

L. Krupička
e-mail: lukas.krupicka@fsv.cvut.cz

R. Štefan
Department of Concrete and Masonry Structures, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic
e-mail: radek.stefan@fsv.cvut.cz

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_9

the conservation equation for total mass of liquid water and ice:

$$\frac{\partial\,(\rho_\ell\,\theta_\ell)}{\partial t}+\frac{\partial\,(\rho_i\,\theta_i)}{\partial t}=\nabla\cdot(\rho_\ell\,\theta_\ell K_h\nabla(h_\ell+z));\tag{1}$$

the energy conservation equation:

$$c_p\frac{\partial\vartheta}{\partial t}+L_f\frac{\partial\,(\rho_i\,\theta_i)}{\partial t}=\nabla\cdot(\lambda\nabla\vartheta)+c_p^\ell\,\rho_\ell\,\theta_\ell\,K_h\nabla(h_\ell+z)\cdot\nabla\vartheta.\tag{2}$$

In (1) and (2), $h_\ell = h_\ell(\mathbf{x},t)$ [m] and $\vartheta = \vartheta(\mathbf{x},t)$ [K] (single-valued functions of the time t and the spatial position $\mathbf{x} \in \Omega$) are the pressure head and temperature, $\theta_\ell = \theta_\ell(h_\ell)$ [-] is the liquid water content, $\theta_i = \theta_i(\vartheta, h_\ell)$ [-] is the ice water content, $K_h = K_h(h_\ell)$ [m s^{-1}] represents the hydraulic conductivity, $c_p = c_p(\vartheta, h_\ell)$ [J m^{-3} K^{-1}] is the effective heat capacity and $\lambda = \lambda(\vartheta, h_\ell)$ [W m^{-1} K^{-1}] is the thermal conductivity. Material constant parameters are the volumetric heat capacity of water c_p^ℓ (4.181×10^6 Jm^{-3} K^{-1}), the density of liquid water ρ_ℓ (approximately 1000.0 kg m^{-3}), the density of ice ρ_i (918 kg m^{-3}), and L_f is the latent heat of fusion (3.34×10^5 J kg^{-1}).

2 Constitutive and Thermodynamic Relationships

Water in pores does not freeze at 273.15 K, but is subject to a freezing-point depression caused by interaction between water, particles, and solutes. The generalized Clapeyron equation is used to describe the condition for the coexistence of water and ice [6]:

$$\frac{\mathrm{d}p_\ell}{\mathrm{d}\vartheta}=\frac{\rho_\ell L_f}{\vartheta},\tag{3}$$

where p_ℓ is the liquid water pressure [Pa], ρ_ℓ the density of liquid water (approximately 1000.0 kg m^{-3}), and L_f is the latent heat of fusion (3.34×10^5 J kg^{-1}). Define h_ℓ [m], $h_\ell\rho_\ell g = p_\ell$, as the matric potential corresponding to the liquid water content θ_ℓ [-] and the matric potential h_w [m], $h_w\rho_\ell g = p_w$, corresponding to the total water content θ_w [-] (liquid and ice). Here, g is the acceleration due to gravity (9.81 m s^{-2}).

Let $\vartheta_0 = 273.15$ K be the freezing temperature at the saturation pressure $p_\ell = 0$ Pa. If the porous media is unsaturated (at the pressure $p_\ell < 0$), the surface tension at the water/air interface decreases the water freezing/melting temperature to $\vartheta_{f/m} <$ 273.15 K. When $\vartheta \geqslant \vartheta_{f/m}$, all water is unfrozen. When $\vartheta < \vartheta_{f/m}$, the porous media is under freezing conditions and the liquid water pressure p_ℓ depends on the intensity of freezing condition provided by ϑ. Denote by p_w the total water pressure corresponding to the total water content (liquid and ice). Above the freezing temperature $\vartheta_{f/m}$ all of ice melted and the total water pressure p_w and the liquid water pressure p_ℓ coincide. Consequently, solving (3) we deduce that for the total water pressure p_w the freezing/melting temperature corresponds to

$$\vartheta_{f/m} = \vartheta_0 \exp\left(\frac{p_w}{\rho_\ell L_f}\right) = \vartheta_0 \exp\left(\frac{h_w g}{L_f}\right) \approx \vartheta_0 + \frac{g\vartheta_0}{L_f} h_w. \tag{4}$$

The formulation of the liquid water matric potential can be determined from (3) using the Heaviside function as

$$\begin{aligned} h_\ell &= h_w + \left(\frac{L_f}{g} \ln\left|\frac{\vartheta}{\vartheta_0}\right| - h_w\right) H(\vartheta_{f/m} - \vartheta) \\ &\approx h_w + \underbrace{\left(\frac{L_f}{g}\frac{\vartheta - \vartheta_0}{\vartheta_0} - h_w\right) H(\vartheta_{f/m}(h_w) - \vartheta)}_{\psi(h_w,\vartheta)} = h_w + \psi(h_w, \vartheta). \end{aligned} \tag{5}$$

The amount of water present at a certain matric potential of the porous medium is characterized by the water retention curve $\Theta(\cdot)$. In particular, $\theta_w = \Theta(h_w)$ and $\theta_\ell = \Theta(h_\ell)$. Here, we use the relation proposed by van Genuchten [5] $\Theta(\xi) = \theta_r + (\theta_s - \theta_r)\left[1 + |\alpha\xi|^n\right]^{-m}$, where θ_s is the saturated water content [-], θ_s is the residual water content [-], α [m^{-1}], m and n are parameters. When $\vartheta \geqslant \vartheta_{f/m}$ all water is unfrozen and, taking into account (5), $h_w = h_\ell$ and $\theta_w = \theta_\ell$. Whenever $\vartheta < \vartheta_{f/m}$, the ice fraction θ_i [-] can be expressed as $\theta_i = \theta_w - \theta_\ell$ [-]. The total water content θ_M as derived by the fraction of total mass of liquid water and ice reads $\theta_M = \theta_\ell + \frac{\rho_i}{\rho_\ell}\theta_i$.

3 Complete Mathematical Model

Let $T > 0$ be the fixed value and Ω be the Lipschitz domain in $\mathbb{R}^N$, $N = 1, 2, 3$, with boundary Γ. Denote $I = (0, T)$, $\Omega_T = \Omega \times I$, and $\Gamma_T = \Gamma \times I$. The mathematical model consists of the following initial boundary value problem:

$$\frac{\partial \theta_M}{\partial t} = \nabla \cdot \left(K_h \nabla \left(h_w + z + \psi\right)\right) \qquad in\ \Omega_T, \tag{6}$$

$$c_a \frac{\partial \vartheta}{\partial t} + L_f \rho_\ell \frac{\partial \theta_i}{\partial t} = \nabla \cdot (\lambda \nabla \vartheta) + c_p^\ell\, \rho_\ell\, K_h \nabla(h_w + z + \psi) \cdot \nabla \vartheta \qquad in\ \Omega_T, \tag{7}$$

$$-K_h \nabla(h_w + \psi) \cdot \mathbf{n} = q_\ell, \quad -\lambda \nabla \vartheta \cdot \mathbf{n} = \alpha_c(\vartheta - \vartheta_\infty) + q_H \qquad in\ \Gamma_T, \tag{8}$$

$$h_w = (h_w)_0, \quad \vartheta = \vartheta_0 \qquad in\ \Omega. \tag{9}$$

This system describes the coupled water flow and heat transport involving freezing–thawing processes in porous media. Equations (6) and (7) represent conservation laws for mass and energy, the Eq. (8) prescribes boundary conditions of Neumann type and the Eq. (9) represents appropriate initial conditions. In (6–9), h_w and ϑ are the primary unknowns. Further, $c_a = c_p - \frac{\rho_\ell}{\rho_i}\frac{d\theta_i}{d\vartheta}$ [J m^{-3} K^{-1}] is the so-called apparent heat capacity and q_ℓ, q_H, ϑ_∞, $(h_w)_0$, and ϑ_0 are given smooth functions.

4 Structural Conditions and Assumptions on Physical Parameters

Let us present some properties and additional assumptions on physical parameters introduced in the model.

(A1) The parameters $\rho_\ell, \rho_i, \theta_s, \theta_r, c_p^\ell, L_f$, and α_c are real positive constants, $\rho_i < \rho_w$.

(A2) The thermal conductivity λ, apparent thermal capacity c_a, and hydraulic conductivity K_h are assumed to be positive continuous functions of their arguments (see [2] for specific examples). In addition, $0 < c_a \leqslant c_a^\sharp < +\infty\,(c_a^\sharp = \text{const} > 0)$.

(A3) $\Theta(\cdot)$ is positive, nondecreasing, continuous, and bounded function such that $\theta_r \leqslant \Theta(\xi) \leqslant \theta_s\ \forall\xi \in \mathbb{R}$.

Consequently, θ_M is a positive continuous function such that

$$0 < \theta_M(\xi,\zeta) = \frac{\rho_i}{\rho_w}\theta_w(\xi) + \left(1 - \frac{\rho_i}{\rho_\ell}\right)\theta_\ell(\zeta) \leqslant \theta_s \quad \textit{for all } \xi,\zeta \in \mathbb{R}.$$

5 The Approximate Solution

Albeit the coupled problem (6–9) is essentially nonstationary in its nature, we shall formulate and analyze a weak form of the stationary problem. It has a significant mathematical interest because the time discretization of the evolution problem leads, in each time step, to a coupled system of stationary equations. Let $0 = t_0 < t_1 < \cdots < t_N = T$ be an equidistant partitioning of time interval $[0;T]$ with step Δt. Set a fixed integer n such that $0 \leqslant n \leqslant N-1$. In what follows, we abbreviate $\phi(z,t_n)$ by ϕ^n ($\equiv \phi(z)^n$) for any function ϕ. The time discretization of the continuous model is accomplished through a semi-implicit difference scheme. Consequently, we have to solve, successively for $n = 0,\ldots,N-1$, the following semi-linear system with primary unknowns $[\vartheta^{n+1}, h_w^{n+1}]$

$$\frac{\theta_M^{n+1} - \theta_M^n}{\Delta t} = \nabla\cdot\left(K_h^n \nabla h_w^{n+1}\right) + \nabla\cdot\left(K_h^n\nabla(\psi^n + z)\right), \tag{10}$$

$$c_a^n\frac{\vartheta^{n+1} - \vartheta^n}{\Delta t} + L_f\rho_\ell\frac{\theta_M^{n+1} - \theta_M^n}{\Delta t} = \nabla\cdot\left(\lambda^n\nabla\vartheta^{n+1}\right) + c_p^\ell\rho_\ell K_h^n\nabla(h_w^n + z + \psi^n)\cdot\nabla\vartheta^n, \tag{11}$$

$$-K_h^n\nabla(h_w^{n+1} + \psi^n)\cdot\mathbf{n} = q_\ell^{n+1} on \Gamma, \tag{12}$$

$$-\lambda^n\nabla\vartheta^{n+1}\cdot\mathbf{n} = \alpha_c(\vartheta^{n+1} - \vartheta_\infty^{n+1}) + q_H^{n+1} on \Gamma \tag{13}$$

Here, we assume that the functions h_w^n and ϑ^n are known and we put $K_h^n = K_h(\vartheta^n, h_w^n)$, $\lambda^n = \lambda(\vartheta^n, h_w^n)$, $c_a^n = c_a(\vartheta^n, h_w^n)$. In what follows, we study the problem of the existence of the variational solution ϑ^{n+1} and h_w^{n+1}: to find the couple

$[\vartheta^{n+1}, h_w^{n+1}] \in W^{1,r}(\Omega)^2$, $r > 2$, such that

$$\frac{1}{\Delta t}\int_\Omega \left(\theta_M^{n+1} - \theta_M^n\right)\phi_1 + c_a^n\left(\vartheta^{n+1} - \vartheta^n\right)\phi_2 + L_f\rho_\ell\left(\theta_M^{n+1} - \theta_M^n\right)\phi_2 \mathrm{d}\Omega$$
$$+\int_\Omega K_h^n\nabla h_w^{n+1}\cdot\nabla\phi_1\mathrm{d}\Omega + \int_\Omega K_h^n\nabla\psi^n\cdot\nabla\phi_1\mathrm{d}\Omega + \int_\Omega \nabla K_h^n\cdot\mathbf{e}_z\,\phi_1\mathrm{d}\Omega$$
$$+\int_\Omega \lambda^n\nabla\vartheta^{n+1}\cdot\nabla\phi_2\mathrm{d}\Omega - \int_\Omega c_p^\ell\rho_\ell K_h^n\nabla(h_w^n + z + \psi^n)\cdot\nabla\vartheta^n\,\phi_2\,\mathrm{d}\Omega$$
$$+\int_\Gamma q_\ell^{n+1}\phi_1\,\mathrm{d}S + \int_\Gamma \alpha_c(\vartheta^{n+1} - \vartheta_\infty^{n+1})\phi_2 + q_H^{n+1}\phi_2\,\mathrm{d}S = 0 \quad (14)$$

holds for every $[\phi_1, \phi_2] \in W^{1,r'}(\Omega)^2$, $r' = r/(r-1)$, and

$$\vartheta^0(\mathbf{x}) = \vartheta_0(\mathbf{x}) \text{ and } h_w^0(\mathbf{x}) = (h_w)_0(\mathbf{x}) \text{ in } \Omega.$$

Theorem 1 *Assume that* $[h_w^n, \vartheta^n] \in W^{1,s}(\Omega)^2$ *with some* $s > 2$ *is known and let the assumptions* (A1–A3) *be satisfied. Then, there exists the variational solution* $[\vartheta^{n+1}, h_w^{n+1}] \in W^{1,r}(\Omega)^2$ *with some* $r > 2$*, of the problem* (10–13).

Remark 1 (Remark to the proof) The problem can be associated with the operator equation $\mathcal{A}([\vartheta^{n+1}, h_w^{n+1}]) = \mathbf{f}$. It can be shown that the operator $\mathcal{A} : W^{1,2}(\Omega)^2 \to [W^{1,2}(\Omega)^2]^*$ is pseudomonotone and coercive. Now [4, Theorem 3.3.42] yields the existence of the solution $[\vartheta^{n+1}, h_w^{n+1}] \in W^{1,2}(\Omega)^2$ to the equation $\mathcal{A}([\vartheta^{n+1}, h_w^{n+1}]) = \mathbf{f}$ for every $\mathbf{f} \in [W^{1,2}(\Omega)^2]^*$. It can be shown that for $[h_w^n, \vartheta^n] \in W^{1,s}(\Omega)^2$ with some $s > 2$, we have $\mathbf{f} \in [W^{1,s'}(\Omega)^2]^*$, $s' = s/(s-1)$. Now the regularity of the solution $[\vartheta^{n+1}, h_w^{n+1}] \in W^{1,r}(\Omega)^2$ with some $r > 2$ follows from [1, Theorem 2].

6 Example

By means of the model described above, we briefly present the numerical simulation of benchmark experiment in [3]. The soil thickness in the numerical simulation for the one-dimensional vertical transport is 0.2 m. The initial uniform temperature is set to 279.85 K and the uniform water content to 0.33. The top of the column is exposed to the temperature 267.15 K; hence, it is subjected to freezing from top to down. All boundaries are hydraulic insulated. Physical properties of soil are taken from [2, 5] and the basic material constants are summarized in Table 1. The spatial discretization of the system (10–13) is carried out by means of the finite element method. This resulting system is solved using the well-known Newton method at each time step with $\Delta t = 1$ s. The progress of water and ice content and temperature at 12 h based on numerical simulation is shown in Figs. 1 and 2. Simulated results demonstrate a trend that water moves toward the freezing front as commonly observed by experimental phenomena (see [3]). The total water content in the frozen region increases and liquid

Table 1 Constant parameters

Constant	Value	Dimension	Description
ρ_ℓ	1000.0	kg m^{-3}	Density of liquid water
ρ_i	918.0	kg m^{-3}	Density of ice
θ_s	0.535	m^3 m^{-3}	Saturated water content
θ_r	0.05	m^3 m^{-3}	Residual water content
c_p^ℓ	4.181×10^6	J m^{-3} K^{-1}	Volumetric heat capacity of liquid water
L_f	3.34×10^5	J kg^{-1}	Latent heat of fusion
α_c	28.0	W m^{-2} K^{-2}	Convection heat transfer coefficient

Fig. 1 Distribution of water and ice at 12 h

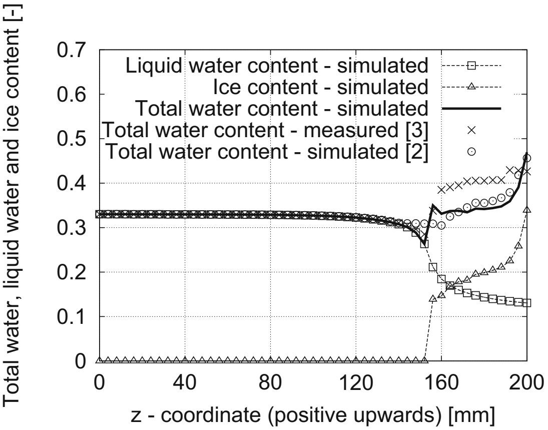
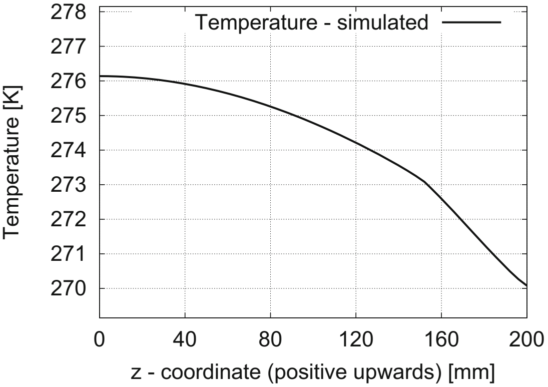


Fig. 2 Distribution of temperature at 12 h



water from lower region moves upward and increases the contribution of ice at the upper surface.

Acknowledgments This research was supported by the project GAČR 13-18652S (M. Beneš). In addition, this work was partly covered by the grant SGS13/001/OHK1/1T/11 provided by the Grant Agency of the Czech Technical University in Prague (L. Krupička).

References

1. Gallouët, T., Monier, A.: On the regularity of solutions to elliptic equations. Rendiconti di Matematica, Serie VII **19**, 471–488 (1999)
2. Liu, Z., Yu, X.: Coupled thermo-hydro-mechanical model for porous materials under frost action: theory and implementation. Acta Geotech. **6**, 51–65 (2011)
3. Mizoguchi, M.: Water, heat and salt transport in freezing soil. Ph.D. thesis. (In Japanese.) University of Tokyo (1990)
4. Nečas, J.: Introduction to the theory of nonlinear elliptic equations. Teubner-Texte zur Mathematik. Leipzig (1983)
5. van Genuchten, M.T.: A closed form equation for predicting the hydraulic conductivity of unsaturated soil. Soil Sci. Soc. Am. J. **44**, 892–898 (1980)
6. Williams, P., Smith, M.: The Frozen Earth: Fundamentals of Geocryology. Cambridge University Press, Cambridge (1989)

Set-valued Nonlinear Fredholm Integral Equations: Direct and Inverse Problem

M.I. Berenguer, H. Kunze, D. La Torre and M. Ruiz Galán

Abstract In this chapter we study a set-valued nonlinear Fredholm integral inclusion. We prove the existence of a solution and provide a numerical method based on the Steiner selection and Schauder bases to determine an approximated solution. We then discuss an inverse problem. Numerical results are also provided to show how the method works practically.

1 Set-valued Fredholm Integral Equation

We consider the following set-valued Fredholm integral equation

$$x(t) \in f_0(t) + \int_0^1 F(t, s, x(s))ds \tag{1}$$

where $f_0 : [0, 1] \rightrightarrows \mathbb{R}^N$ and $F : [0, 1] \times [0, 1] \times \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ are set-valued mappings, and $x : [0, 1] \to \mathbb{R}^N$ is the unknown solution that has to be determined. Such integral inclusions arise in modeling systems for which we have no complete description.

In order to handle (1), we consider the so-called *Steiner selection* of the involved set-valued mappings. To be more precise, let us recall that for any subset $K \subset \mathbb{R}^N$, the *support function* is defined as $\text{supp}(K, p) := \sup\{<p, x> : x \in K\}$, $(p \in \mathbb{R}^N)$, $(<\cdot, \cdot>$ denotes the usual inner product) and its subdifferential $\partial(\text{supp}(K, p))$ is

M.I. Berenguer (✉) · M. Ruiz Galán
Department of Applied Mathematics, University of Granada, Granada, Spain
e-mail: maribel@ugr.es

M. Ruiz Galán
e-mail: mruizg@ugr.es

H. Kunze
Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada
e-mail: hkunze@uoguelph.ca

D. La Torre
Department of Applied Mathematics and Sciences, Khalifa University,
Abu Dhabi, UAE

Department of Economics, Management and Quantitative Methods,
University of Milan, Milan, Italy
e-mail: davide.latorre@unimi.it

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_10

given by $\partial(\operatorname{supp}(K,p)) = \{x \in K :< p, x >= \operatorname{supp}(K,p)\}$. For any nonempty, compact, and convex subset K of $\mathbb{R}^N$, the *Steiner point* of K is defined by (see [1, § 9.4.1]) $s_N(K) := \frac{1}{vol(B^N)} \int_{B^N} m(\partial(\operatorname{supp}(K,p)))dp$, where $m(\partial(\operatorname{supp}(K,p)))$ denotes the element of minimal norm in $\operatorname{supp}(K,p)$ and $vol(B^N)$ is the measure of the N-dimensional unit ball B^N of $\mathbb{R}^N$, both for the Euclidean norm $\|\cdot\|_2$. We also need the following standard metric concept: if K and L are compact subsets of $\mathbb{R}^N$ then their Hausdorff distance is given by

$$d_H(K,L) := \max\left\{\max_{x\in K}\min_{y\in L}\|x-y\|_2, \max_{y\in L}\min_{x\in K}\|x-y\|_2\right\}.$$

First, we present an existence result for the solution to the set-valued integral (1). Observe that in terms of the integral operator $T : C([0,1],\mathbb{R}^N) \longrightarrow C([0,1],\mathbb{R}^N)$ defined for each $x \in C([0,1],\mathbb{R}^N)$ and $t \in [0,1]$ as

$$Tx(t) := s_N(f_0(t)) + \int_0^1 s_N(F(t,s,x(s)))ds, \tag{2}$$

when F and f_0 are continuous set-valued mappings in the Hausdorff metric taking nonempty, compact, and convex values, the problem

$$x(t) = s_N(f_0(t)) + \int_0^1 s_N(F(t,s,x(s)))ds, \tag{3}$$

is equivalent to finding a fixed point $x \in C([0,1],\mathbb{R}^N)$ of the operator T where $s_N(f_0(t))$ and $s_N(F(t,s,x(s)))$ are the Steiner points in $f_0(t)$ and $F(t,s,x(s))$, respectively. In the next result we derive, under suitable assumptions, the existence of one and only one solution of (3); hence that the set-valued integral (1) admits a continuous solution.

Proposition 1 *Let $f_0 : [0,1] \rightrightarrows \mathbb{R}^N$ and $F : [0,1]\times[0,1]\times\mathbb{R}^N \rightrightarrows \mathbb{R}^N$ be continuous set-valued mappings in the Hausdorff metric taking nonempty, compact, and convex values. Let us also assume that F is $c-$Lipschitz in its third variable w.r.t. the Hausdorff distance, i.e., $t,s \in [0,1]$, $y_1, y_2 \in \mathbb{R}^N \Rightarrow d_H(F(t,s,y_1), F(t,s,y_2)) \le c\|y_1 - y_2\|_2$, and that $c_N := c(N)^{3/2} < 1$. Then the integral operator T defined in (2) admits a unique solution $\hat{x} \in C([0,1],\mathbb{R}^N)$. Moreover, for any function $x \in C([0,1],\mathbb{R}^N)$ and any $j \ge 1$, $\|T^j x - \hat{x}\| \le \frac{c_N^j}{1-c_N}\|Tx - x\|$.*

Proof According to Banach's fixed point theorem, it suffices to prove that the operator T is c_N-contractive, when endowing the space $C([0,1],\mathbb{R}^N)$ with its sup-sup norm $\|\cdot\|$. So, let $x, y \in C([0,1],\mathbb{R}^N)$. Then, given $t,s \in [0,1]$, the chain of inequalities

$$\begin{aligned}
\|Tx(t) - Ty(t)\|_\infty &\le \int_0^1 \|s_N(F(t,s,x(s))) - s_N(F(t,s,y(s)))\|_\infty ds \\
&\le \int_0^1 \|s_N(F(t,s,x(s))) - s_N(F(t,s,y(s)))\|_2 ds \\
&\le N \int_0^1 d_H(F(t,s,x(s)), \\
&\quad F(t,s,y(s)))ds \;\; \text{(by [1, Theorem 9.4.1])} \\
&\le cN \int_0^1 \|x(s) - y(s)\|_2 ds \;\; \text{(by the lipschitzianity of } F\text{)} \\
&\le c(N)^{3/2} \int_0^1 \|x(s) - y(s)\|_\infty ds \;\le c(N)^{3/2}\|x - y\|
\end{aligned}$$

clearly implies $\|Tx - Ty\| \leq c_N \|x - y\|$, and the proof is complete.

In order to develop an algorithm to approximate the unique solution of (3), and then for the integral inclusion (1), we turn to a certain kind of biorthogonal system, more precisely Schauder bases, in the Banach spaces $C([0,1]) = C([0,1], \mathbb{R})$ and $C([0,1]^2) = C([0,1]^2, \mathbb{R})$, equipped with their usual sup norms, in the next section.

2 Numerical Solution via Schauder Bases

Let us recall that if E is a real Banach space with topological dual space E^*, a family $\{(x_i, x_i^*)\}_{i\in\mathbb{N}}$ in $E \times E^*$ is said to be a *Schauder basis* for E provided that for all $x \in E$, there is a unique sequence $\{\lambda_i\}_{i\geq 1}$ of real numbers such that $x = \sum_{i\geq 1} \lambda_i x_i$. The ith *biorthogonal functional* x_i^* is given at such an x by $x_i^*(x) = \lambda_i$ and the corresponding ith *projection* Π_i by $\Pi_i(x) = \sum_{j=1}^{i} \lambda_j x_j$. In particular, for all $x \in E$ we have that

$$\lim_{i\geq 1} \|\Pi_i x - x\| = 0. \tag{4}$$

It is a well-known fact that, as a consequence of the open mapping theorem, the biorthogonal functionals and the projections are linear and continuous (see for instance [6, Theorem 3.1]). Since the biorthogonal functionals of a Schauder basis are completely determined by the sequence $\{x_i\}_{i\geq 1}$, for simplicity the notation $\{x_i\}_{i\geq 1}$ is often used instead of $\{(x_i, x_i^*)\}_{i\geq 1}$.

For the Banach spaces $C([0,1])$ and $C([0,1]^2)$ the so-called *usual bases* can be considered: let $\{t_i\}_{i\geq 1}$ be a sequence of distinct points in $[0,1]$ such that $t_1 = 0$ and $t_2 = 1$. We define the Schauder basis $\{b_i\}_{i\geq 1}$ of $C([0,1])$ as $b_1(t) := 1$, $t \in [0,1]$, and for $i \geq 1$, b_i is the piecewise linear continuous function on $[0,1]$ with nodes at $\{t_j : 1 \leq j \leq i\}$, given by $b_i(t_i) = 1$ and $b_i(t_j) = 0$ for $j < i$. We denote by $\{b_i^*\}_{i\geq 1}$ and $\{P_i\}_{i\geq 1}$, respectively, their sequences of biorthogonal functionals and projections. The Schauder basis $\{B_i\}_{i\geq 1}$ in $C([0,1]^2)$ is the corresponding bivariate tensor basis of $\{b_i\}_{i\geq 1}$ ([8] and [11]): if for a real number a, $[a]$ denotes its integer part, and $\sigma : \mathbb{N} \longrightarrow \mathbb{N}\times\mathbb{N}$ is the bijective mapping given by

$$\sigma(i) := \begin{cases} (\sqrt{i}, \sqrt{i}), & \text{if } \ [\sqrt{i}] = \sqrt{i} \\ (i - [\sqrt{i}]^2, [\sqrt{i}] + 1), & \text{if } \ 0 < i - [\sqrt{i}]^2 \leq [\sqrt{i}], \\ ([\sqrt{i}] + 1, i - [\sqrt{i}]^2 - [\sqrt{i}]), & \text{if } \ [\sqrt{i}] < i - [\sqrt{i}]^2 \end{cases}$$

then $B_i(t,s) := b_p(t) b_q(s)$, $t, s \in [0,1]$, whenever $\sigma(i) = (p,q)$. $\{B_i^*\}_{i\geq 1}$ and $\{Q_i\}_{i\geq 1}$ stand for the respective sequences of biorthogonal functionals and projections.

The iterative method derived from Proposition 1 for calculating the unique solution $\hat{x}$ of the Eq. (3), equivalently, the unique fixed point of T in (2), presents an obvious limitation if we want to implement it: given $\widehat{x} \in C([0,1], \mathbb{R}^N)$, the calculations that lead to determining each iteration $T^m x$ are not possible, in general, in an explicit way. In order to avoid this disadvantage, we emphasize a certain property of the usual Schauder basis of the Banach space $C([0,1]^2)$. Their projections allow us to discretize each of the mentioned iterations, deriving an approximation of them and making feasible their calculations, in view of the specific form of the operator T and the following property ([8, 11]), which is remarkable from a computational view point: for all $z \in C([0,1]^2)$ we have that $B_1^*(z) = z(t_1, t_1)$ and

$$\left.\begin{array}{c} i \geq 2 \\ \sigma(i)=(p,q) \end{array}\right\} \Rightarrow B_i^*(z) = z(t_p, t_q) - \sum_{j=1}^{i-1} B_j^*(z) B_j(t_p, t_q).$$

Let us now introduce the numerical method for approximating the solution of (3). Starting from $x \in C([0,1], \mathbb{R}^N)$ and $i_1, i_2, \cdots$, we define recursively the continuous functions for $j \in \mathbb{N}$ and $t \in [0,1]$,

$$x_j(t) := s_N(f_0(t)) + \left[\int_0^1 Q_{i_j}(\Psi_{j-1,1}(t,s))ds, \ldots, \int_0^1 Q_{i_j}(\Psi_{j-1,N}(t,s))ds\right]^T, \quad (5)$$

where $x_0(t) := x(t)$, $(t \in [0,1])$, and, for $t, s \in [0,1]$,

$$\Psi_{j-1}(t,s) = s_N(F(t,s,x_{j-1}(s))) = [\Psi_{j-1,1}(t,s), \ldots, \Psi_{j-1,N}(t,s)]^T. \quad (6)$$

Let us consider the simple but intrinsic bounding $\|x_j - \hat{x}\| \leq \|T^j x - x_j\| + \|T^j x - \hat{x}\|$, whose second right hand term has been controlled in Proposition 1. For the first one we have the inequality below:

Proposition 2 *Let $f_0 : [0,1] \rightrightarrows \mathbb{R}^N$ and $F : [0,1] \times [0,1] \times \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ be set-valued mappings satisfying the assumptions in Proposition 1, let $c_N := c(N)^{3/2}$ and let $\hat{x}$ be the unique fixed point of the integral operator T given in 2. Then the inequality*
$\|T^j x - x_j\| \leq \sum_{k=1}^j c_N^{j-k} \|T x_{k-1} - x_k\|$ is valid whenever $x \in C([0,1], \mathbb{R}^N)$.

Proof If $x \in C([0,1], \mathbb{R}^N)$ and $j \geq 1$, then

$$\begin{aligned}
\|T^j x - x_j\| &\leq \|T^j x - T x_{j-1}\| + \|T x_{j-1} - x_j\| \\
&\quad \leq c_N \|T^{j-1} x - x_{j-1}\| + \|T x_{j-1} - x_j\| \\
&\leq c_N \left(\|T^{j-1} x - T x_{j-2}\| + \|T x_{j-2} - x_{j-1}\|\right) + \|T x_{j-1} - x_j\| \\
&\leq c_N^2 \|T^{j-2} x - x_{j-2}\| + c_N \|T x_{j-2} - x_{j-1}\| + \|T x_{j-1} - x_j\| \\
&\leq c_N^2 \left(\|T^{j-2} x - T x_{j-3}\| + \|T x_{j-3} - x_{j-2}\|\right) + c_N \|T x_{j-2} - x_{j-1}\| \\
&\quad\quad + \|T x_{j-1} - x_j\| \leq \cdots \\
&\leq \textstyle\sum_{k=1}^j c_N^{j-k} \|T x_{k-1} - x_k\|, \text{ and we are done.}
\end{aligned}$$

Now we can show that for a suitable choice of $j \geq 1$ and $i_1, i_2, \cdots, i_j$, x_j is as close as desired to the fixed point $\hat{x}$ of the operator T:

Theorem 1 *Let $f_0 : [0,1] \rightrightarrows \mathbb{R}^N$ and $F : [0,1] \times [0,1] \times \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ satisfy the assumptions in Proposition 1 and let $\hat{x}$ be the unique solution of (3). Then for each $\varepsilon > 0$ and each $x \in C([0,1], \mathbb{R}^N)$, there exist $j \geq 1$ and $i_1, i_2, \cdots < i_j$ such that $\|x_j - \hat{x}\| < \varepsilon$, where x_j is the approximate function defined by (5) and (6).*

Table 1 Absolute errors for Example 1

	$i_j = 9$ for $j = 1, \ldots, 6$	$i_j = 17$ for $j = 1, \ldots, 6$	$i_j = 33$ for $j = 1, \ldots, 6$
t	$\lvert x_6(t) - \hat{x}(t)\rvert$	$\lvert x_6(t) - \hat{x}(t)\rvert$	$\lvert x_6(t) - \hat{x}(t)\rvert$
0.	1.5×10^{-3}	3.9×10^{-4}	9.9×10^{-5}
0.2	1.9×10^{-3}	4.6×10^{-4}	1.2×10^{-4}
0.4	1.8×10^{-3}	4.9×10^{-4}	1.1×10^{-4}
0.6	1.8×10^{-3}	4.9×10^{-4}	1.1×10^{-4}
0.8	1.9×10^{-3}	4.6×10^{-4}	1.2×10^{-4}
1	1.5×10^{-3}	3.9×10^{-4}	9.9×10^{-5}

Proof For each $\varepsilon > 0$, we first choose $j \geq 1$ with $\frac{c_N^j}{1-c_N}\|Tx - x\| < \frac{\varepsilon}{2}$, since $c_N < 1$. On the other hand, the convergence property (4) for $\{B_i\}_{i\geq 1}$ guarantees the existence of $i_1, i_2, \cdots, i_j$ such that $\sum_{k=1}^{j} c_N^{j-k}\|Tx_{k-1} - x_k\| < \frac{\varepsilon}{2}$. Finally, it follows from the two bounds and Propositions 1 and 2 that $\|x_j - \hat{x}\| \leq \|T^j x - x_j\| + \|T^j x - \hat{x}\| < \varepsilon$.

Example 1 Let us consider the set-valued Fredholm integral equation
$x(t) \in \left[-\frac{59}{20} + 2t + \frac{9}{5}t^2 + \frac{2}{3}t^3, \frac{5}{2}\right] + \int_0^1 \left[\frac{1}{5}s^2 + \frac{1}{10}x(s), \frac{1}{5}t^2 + \frac{3}{5}s^2 + \frac{1}{10}x(s)\right] ds$
where $\hat{x}(t) = \frac{s^3}{3} + s^2 + s$ is the unique fixed point of the associate operator T defined in (2). In order to construct the Schauder basis $\{B_i\}_{i\geq 1}$ in $C([0,1]^2, \mathbb{R})$, we consider $t_1 = 0$, $t_2 = 1$ and for $n \in \mathbb{N} \cup \{0\}$, $t_{j+1} = \frac{2k+1}{2^{n+1}}$ if $j = 2^n + k + 1$ where $0 \leq k < 2^n$ are integers. To define the sequence $\{x_j\}_{j\geq 1}$, we take $x_0(t) = s_N(f_0(t))$. We include the results when we approximate the unique solution of the operator T associated defined by (2) by the forth iteration, taking $i_j = 9, 17$ or 33 for $j = 1, 2, 3$ and 4. The algorithms associated with the numerical method were performed using Mathematica 7. The results are shown in Table 1.

It is worth mentioning that the numerical method introduced above extends the scalar and single-valued case developed in [2], and that Schauder bases have been successfully used in the numerical study of different integral, differential, and integro-differential problems (see [3–5, 7]).

3 Inverse Problem

An inverse problem consists of estimating the values of unknown parameters in a model by using an empirical target solution. Kunze et al. [9] present a framework for solving the inverse problem of estimating f in the Fredholm integral equation (1) given x; we aim to extending it to the set-valued Fredholm integral inclusion. Starting from a target x, and a parameter-dependent family of set-valued integral Fredholm operators taking the form

$$T_\lambda : x \in C([0,1]) \to f_0(t) + \int_0^1 F_\lambda(t, s, x(s))ds \tag{7}$$

we seek a parameter value $\lambda \in \Lambda \subset \mathbb{R}^s$ such that $d(x, T_\lambda x)$ is small enough, where d is the distance point-to-set. Let us notice that the hypotheses on all F_λ imply that the function $t \to f_0(t) + \int_0^1 F_\lambda(t,s,x(s))ds$ takes compact and closed values. Proposition 3 states a collage-type result for problem (7). We need the following Lemma.

Lemma 1 *[10] Let $a, b \in \mathbb{R}^n$, and $C, D \subset \mathbb{R}^n$ be two compact sets. Then $\|a-b\| \le d(a,C) + d(b,D) + d_H(C,D)$, where $d(a,C)$ is the distance point-to-set and d_H is the Hausdorff distance.*

Proposition 3 *Let $f_0 : [0,1] \rightrightarrows \mathbb{R}^N$ and $F_\lambda : [0,1] \times [0,1] \times \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ be continuous set-valued mappings in the Hausdorff metric taking nonempty, compact, and convex values. Let us also assume that F_λ is a contraction with contractivity factor $c_\lambda < 1$ in its third variable w.r.t. the Hausdorff distance, i.e., $t, s \in [0,1],\ y_1, y_2 \in \mathbb{R}^N \Rightarrow d_H(F_\lambda(t,s,y_1), F_\lambda(t,s,y_2)) \le c_\lambda \|y_1 - y_2\|$, and that $(c_\lambda)_N := c_\lambda (N)^{3/2} < 1$. Let $\hat{x}_\lambda$ be a solution to the following Fredholm integral inclusion $\hat{x}_\lambda(t) \in f_0(t) + \int_0^1 F_\lambda(t,s,\hat{x}_\lambda(s))ds$ for all $t \in [0,1]$, and x be a continuous target. Then $\|x - \hat{x}_\lambda\|_\infty \le \frac{1}{1-c_\lambda} \sup_{t\in[0,1]} d_H \left(x(t), f_0(t) + \int_0^1 F_\lambda(t,s,x(s))ds\right)$*

Proof By computing we have for all $t \in [0,1]$,

$$\begin{aligned}\|\hat{x}_\lambda(t) - x(t)\| &\le d(\hat{x}_\lambda(t), T_\lambda \hat{x}_\lambda(t)) + d(x(t), T_\lambda x(t)) + d_H(T_\lambda x_\lambda(t), T_\lambda x(t))\\ &\le 0 + d(x(t), T_\lambda x(t)) + c_\lambda \|x - \hat{x}_\lambda\|_\infty\\ &\le d_H(x(t), T_\lambda x(t)) + c_\lambda \|x - \hat{x}_\lambda\|_\infty,\end{aligned}$$

which implies the thesis. The last inequality comes from the following calculations:

$$\begin{aligned}d_H(T_\lambda x_\lambda(t), T_\lambda x(t)) &= \sup_{p\in S^1} |\text{supp}(p, T_\lambda x_\lambda(t)) - \text{supp}(p, T_\lambda x(t)|\\ &\le \sup_{p\in S^1} \left|\text{supp}\left(p, \int_0^1 F_\lambda\left(t,s,\hat{x}_\lambda(s)\right)ds\right)\right.\\ &\quad \left. -\text{supp}\left(p, \int_0^1 F_\lambda(t,s,x(s))ds\right)\right|\\ &\le \sup_{p\in S^1} \left|\int_0^1 \text{supp}(p, F_\lambda\left(t,s,\hat{x}_\lambda(s)\right)ds)\right.\\ &\quad \left. - \int_0^1 \text{supp}(p, F_\lambda(t,s,x(s))ds\right|\\ &\le \int_0^1 d_H\left(F_\lambda\left(t,s,\hat{x}_\lambda(s)\right), F_\lambda(t,s,x(s))\right) \le c_\lambda \|\hat{x}_\lambda - x\|_\infty,\end{aligned}$$

where supp(p, S) is the support function in the direction p.
Given a family of set-valued mappings F_λ, $\lambda \in \Lambda$, each satisfying the hypotheses of Proposition 3, provided all c_λ are bounded away from 1, the Proposition allows us to control the error in approximating the x by the inclusion solution $\hat{x}_\lambda$ via minimizing the collage distance portion of the right hand side of (1) W.r.t. $\lambda \in \Lambda$. In the case that the set-valued functions F_λ take compact interval-values $[(F_\lambda)_{\min}, (F_\lambda)_{\max}]$, this calculation reduces to the minimization of the collage distances between the target

x and integral operators which just depend on $(F_\lambda)_{\min}$ and $(F_\lambda)_{\max}$. We illustrate this observation in the following example.

Acknowledgement This work was supported by the GENIL Scholarship of the University of Granada.

References

1. Aubin, J.P., Frankowska, H.: Set-valued Analysis. Birkḧauser, Boston (1990).
2. Berenguer, M.I., Fernández Muñoz, M.V., Garralda-Guillem, A.I., Ruiz Galán M.: Numerical treatment of fixed point applied to the nonlinear Fredholm integral equation. Fix. Point Theory Appl. **2009,** 8 pp. (2009) (Article ID 735638)
3. Berenguer, M.I., Gámez, D., Garralda-Guillem, A.I., Serrano Pérez, M.C.: Nonlinear Volterra integral equation of the second kind and biorthogonal systems. Abstract Appl. Anal. **2010,** 11 pp. (2010) (Article ID 135216)
4. Berenguer, M.I., Garralda-Guillem, A.I., Ruiz Galán, M.: An approximation method for solving systems of Volterra integro-differential equations. Numer. Math. **67,** 126–135 (2013)
5. Calió, F., Garralda-Guillem, A.I., Marchetti, E., Ruiz Galán, M.: About some numerical approaches for mixed integral equations. Appl. Math. Comput. **219,** 464–474 (2012)
6. Carothers, N.L.: A short course on Banach space theory. Lond. Math. Soc. Student Texts **64**, Cambridge University Press, 2004.
7. Gámez, D., Garralda-Guillem, A.I., Ruiz Galán, M.: High-order nonlinear initial-value problems countably determined. J. Comput. Appl. Math. **228,** 77–82 (2009)
8. Gelbaum, B., Gil de Lamadrid, J.: Bases on tensor products of Banach spaces. Pacific J. Math. **11,** 1281–1286 (1961)
9. Kunze, H.E., La Torre, D., Lever, K.M., Vrscay, E.R.: Solving inverse problems for Hammerstein integral equation and its random analog using the "collage method" for fixed points. Int. J. Pure Appl. Math. **60,** 393–408 (2010)
10. Kunze, H.E., La Torre, D., Lever, K.M., Vrscay, E.R.: Fractal Based Methods in Analysis. Springer, Berlin (2012)
11. Semadeni, Z.: Product Schauder bases and approximation with nodes in spaces of continuous functions. Bull. Acad. Polon. Sci. **11,** 387–391 (1963)

Stabilizing Role of Predators in Niche Construction Modeling

Faina S. Berezovskaya and Georgiy P. Karev

Abstract In this chapter a question of "how much over-consumption a renewable resource can tolerate" is addressed using a mathematical model, where a consumer population competes for the common resource, can contribute to resource restoration, and is subject to attacks of predators. The bifurcation analysis of the system shows that well-adapted predators can keep the system in a stable equilibrium even for "strong" prey over-consumption, when the initial system of resource–consumer goes to extinct. Thus, predators may extend the domain of total model system coexistence in niche.

1 Introduction

Modeling of the predator–prey interaction of populations has long history beginning from classical works of V. Volterra [6] (see also [2, 5] etc.). In this work we consider dynamics of the model where a prey population, being a subject of predator attacks, consumes the renewable resource in such a way that can contribute to resource restoration [3, 4]. The model reads:

$$\begin{cases} \frac{dN}{dt} &= N\left(c - \frac{N}{z} - p\right) \\ \frac{dp}{dt} &= \beta p(N - m) \\ \frac{dz}{dt} &= \gamma - \delta z + \frac{e(1-c)N}{N+z} \end{cases} \tag{1}$$

where N, p, z are normalized densities of prey/consumers, predators, and resource, correspondingly. Parameters of this model are γ, δ, which characterize a natural restoration and decay rates of resource, $m > 0$, $\beta \leq 1$, which are a level of a stable

F. S. Berezovskaya (✉)
Department of Mathematics, Howard University, Washington, DC 20059, USA
e-mail: fberezovskaya@howard.edu

G. P. Karev
National Centre for Biotechnology Information, Bethesda, MD 20894, USA
e-mail: karev@ncbi.nlm.nih.gov

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_11

coexistence of predators and preys and a coefficient of the transformation of a prey to predator biomass, correspondingly. Parameters $c > 0$, $e > 0$, characterize the initial (Malthusian) growth rate of consumers and efficiency of resource restoration. The consumers contribute to restoration of a resource if $c < 1$ and exhaust it if $c > 1$ (they are called over-consumers in the latter case), such that the value $e(1-c)$ is the "order of contribution" of consumers to resource.

2 A Model of Consumers-Renewable Resource

The dynamics of the model "consumers-renewable resource"

$$\begin{cases} \frac{dN}{dt} &= N\left(c - \frac{N}{z}\right) \\ \frac{dz}{dt} &= \gamma - \delta z + \frac{e(1-c)N}{N+z} \end{cases} \tag{2}$$

has been investigated in [3]. For all considering parameters the system has nonnegative equilibria: non-hyperbolic point $O_2(0,0)$, whose structure depends on parameter variation, and saddle point $B_2(N = 0, z = \frac{\gamma}{\delta})$; for $0 \le \frac{\gamma}{e} < \frac{c(c-1)}{c+1}$ the system also has nontrivial topological node $A_2(N = \frac{c}{\delta}(\gamma + \frac{ec(c-1)}{c+1}), z = \frac{1}{\delta}(\gamma + \frac{ec(c-1)}{c+1}))$. System (2) demonstrates a wide range of behaviors when parameters vary. Bifurcation diagram of the system is schematically presented in Fig. 1 in the form of $(c, \frac{\gamma}{e})$-parameter and (N, z)-phase portraits. The bifurcation boundaries are presented in Table 1. As the value of c increases the equilibrium A_2 goes through Domain 1 where it is a globally stable node, then through domains of bistability 2,3, where two equilibria A_2 and origin O_2 are locally stable and share the basins of attraction; A_2 loses stability due to the subcritical Hopf bifurcation when it intersects the boundary between Domains 3 and 4 or by the supercritical Hopf bifurcation when intersects the boundary between Domains 3 and 6. The latter event is accompanied by the appearance of a stable limit cycle in Domain 6. Finally the equilibrium A_2 enters Domain 5 yielding an elliptic sector (a family of homoclinic trajectories such that every trajectory tends to the origin point O_2 as $t \to \pm\infty$).

3 Predator-Induced and Predator-Free Equilibria. Bifurcation Diagram of Model (1) in (c, m)-Plane

System (1) can have up to five equilibrium points. Three of them,

$$O(0,0,0), \quad B\left(N = 0,\ p = 0,\ z = \frac{\gamma}{\delta}\right),$$

$$A\left(N = \frac{c}{\delta}\left(\gamma + \frac{ec(c-1)}{c+1}\right),\ p = 0,\ z = \frac{1}{\delta}\left(\gamma + \frac{ec(c-1)}{c+1}\right)\right),$$

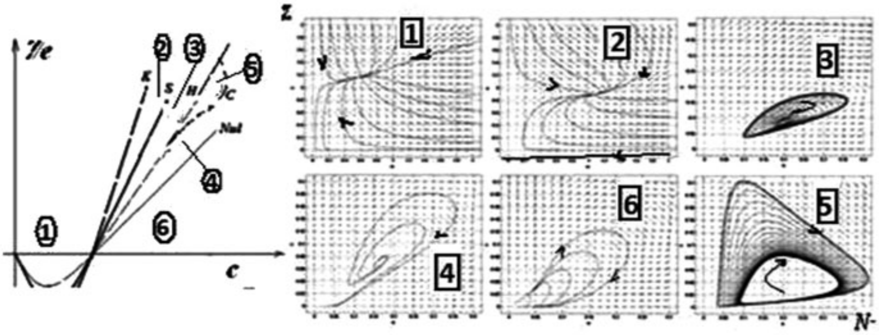


Fig. 1 Bifurcation diagram of system (1) for parameters $(c, \gamma/e)$ and variables (N, z) at fixed $e = 1$ and positive δ. Equilibrium A_2 is globally stable in Domain 1; it shares basins of attraction with equilibrium O_2 in Domains 2 and 3. Only equilibrium O_2 is globally stable in Domains 4 (containing an unstable nontrivial A_2), in Domain 5 O_2 contains an elliptic sector in its vicinity. Domain 6 exists only for certain region of it has stable equilibrium O_2 and stable limit cycle that contains inside unstable equilibrium, their basins are separated by the unstable limit cycle. The boundaries between domains are K, S, H, Nul, C; they correspond, respectively, to the appearance of an attractive sector close to O_2, an unstable limit cycle containing A_2 inside, the change of stability of A_2 via Hopf bifurcations, merging of A_2 and O_2 and saddle-node bifurcation of limit cycles (see Table 1)

Table 1 Domain boundaries of the model bifurcation

Domains	Boundary	Bifurcation
1,2	$\boldsymbol{K}$:$\frac{\boldsymbol{\gamma}}{\boldsymbol{e}} = \boldsymbol{c}_{\mathbf{b}} - 1$	Appearance of stable parabolic sector in a positive neighborhood of equilibrium $\boldsymbol{O}_2$
2,3	$\boldsymbol{S}$:$\boldsymbol{c}(\boldsymbol{\gamma}) = \boldsymbol{c}_{\mathbf{s}}$ no analytical description	Unstable heteroclinics of $\boldsymbol{B}_2$ and $\boldsymbol{O}_2$ separatrixes
3,4 small $\boldsymbol{\delta}$	$\boldsymbol{H}^+ : \frac{\boldsymbol{\gamma}}{\boldsymbol{e}} = \frac{c_{\mathbf{h}}(c_{\mathbf{h}}-1)(c_{\mathbf{h}}(c_{\mathbf{h}}+1)+\delta(c_{\mathbf{h}}+2))}{(c_{\mathbf{h}}+1)^2(c_{\mathbf{h}}+\delta)}$ The first Lyapunov value is positive	Subcritical Hopf bifurcation of equilibrium $\boldsymbol{A}_2$
3,6 big $\boldsymbol{\delta}$	$\boldsymbol{H}^- : \frac{\boldsymbol{\gamma}}{\boldsymbol{e}} = \frac{c_{\mathbf{h}}(c_{\mathbf{h}}-1)(c_{\mathbf{h}}(c_{\mathbf{h}}+1)+\delta(+2))}{(c_{\mathbf{h}}+1)^2(c_{\mathbf{h}}+\delta)}$ The first Lyapunov value is negative	Supercritical Hopf bifurcation of equilibrium $\boldsymbol{A}_2$
6,4	$\boldsymbol{C}$:$\boldsymbol{c}(\boldsymbol{\gamma}) = \boldsymbol{c}_{\boldsymbol{c}}$ no analytical description	Fold bifurcation of limit cycles
4,5	$\boldsymbol{Null} : \frac{\boldsymbol{\gamma}}{\boldsymbol{e}} = \frac{c_0(c_0-1)}{(c_0+1)}$	Merging of equilibria $\boldsymbol{A}_2$ and $\boldsymbol{O}_2$

have the same (N, z)-coordinates as the corresponding points of system (2). In what follows we use the notations $N(A)$, $p(A)$, $z(A)$ for corresponding coordinates of the point A. System (1) can have up to two predator-induced equilibria

$$C^{\pm}\left(N = m, p = c - \frac{m}{z^{\pm}}, z = z^{\pm}\right)$$

with

$$z^{\pm} = \frac{\gamma - \delta m \pm \sqrt{D}}{2\delta}$$

satisfying the equation

$$\delta z^2 - (\gamma - \delta m)z - m(e(1-c)+\gamma) = 0.$$

The domain where $C^{\pm}$-equilibria exist is defined by the condition

$$D = (\gamma + \delta m)^2 + 4e\delta m(1-c) \geq 0;$$

the union of branches $m_{\Delta}^{+} \bigcup m_{\Delta}^{-}$, where

$$m_{\Delta}^{\pm} : m = \frac{2\sqrt{(c-1)e((c-1)e-\gamma)} \pm (c-1)e((c-1)e-\gamma)}{\delta}$$

is the boundary $D = 0$ of this domain. Now define the curve

$$m_{N(A)} : m = \frac{c}{\delta}\left(\gamma + \frac{ec(c-1)}{c+1}\right)$$

with $0 < c < c_0$, where c_0 is a positive root of the equation

$$\frac{\gamma}{e} = \frac{c(c-1)}{c+1}$$

(see Fig. 2 and Table 1). For $c = c_0$ the equilibria A_2 and O_2 of system (2) merge, A_2 leaves the positive quadrant if $c > c_0$ (see Fig. 1, Domain 5). The curve $m_{N(A)}$ and the branch m_{Δ}^{-} of the boundary $D = 0$ have a common point $a\ (c_d, m_d)$ where c_d is a positive root of the equation

$$\frac{\gamma}{e} = \frac{c(c-1)(c+2)}{(c+1)^2}$$

and

$$m_d = \frac{c_d}{\delta}\left(\gamma + \frac{ec_d(c_d-1)}{c_d+1}\right).$$

Let DC_1 be the domain in positive part of (c,m)-plane bounded by the curves $m_{N(A)}$ and the interval of axis c: $0 < c < c_0$. Next, let us define the curve

$$M(c,m) = \left\{ \begin{matrix} m_{N(A)} & c < c < c_d \\ m_{\Delta}^{-} & c > c_d \end{matrix} \right\}.$$

Denote DC the domain in the positive quadrant $(c > 0, m > 0)$ bounded by the curve $M(c,m)$ and the c-axes; let $DC_2 = P \setminus DC_1$, $DA = R_{+}^{2} \setminus DC$ (see Fig. 2).

Let the positive parameters γ, e, β, δ are fixed. The structures of equilibria of the system are described by the following statements.

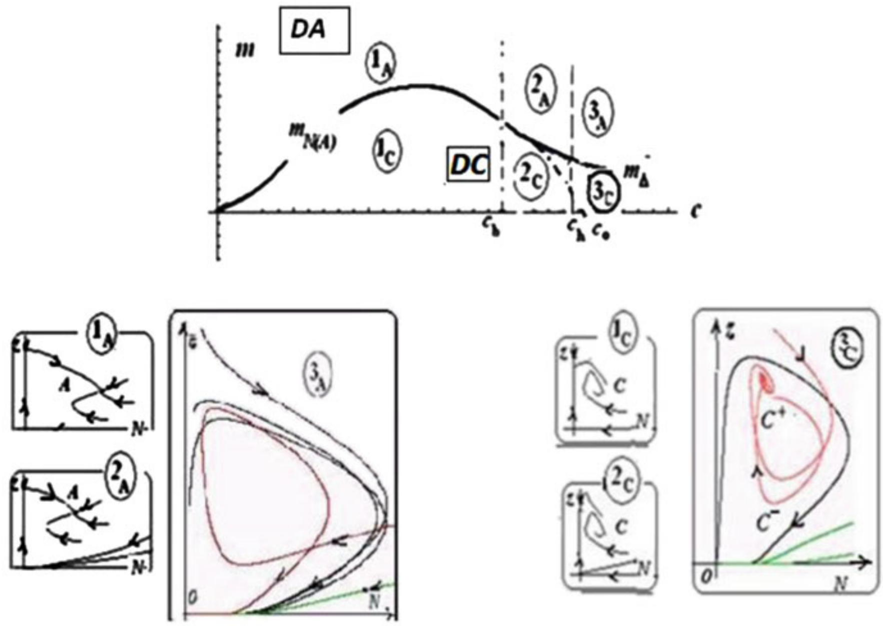


Fig. 2 Schematically presented (N, z)-cut of bifurcation diagram of system (1) in (m, c)-parameter plane. The axis c is partitioned by the points c_b, c_h, c_0, which correspond to bifurcations in model (1; see Table 1 and Fig. 1). The parametric portrait of the system consists of six domains with qualitatively different stable behaviors; the domains are denoted by the integer and subindex A or C, where index A means that equilibrium A is stable and index C means that equilibrium C is stable. The (N, z)-phase portraits of the system are presented in the lower panel. Portraits $3A$ and $3C$ are computed as $\beta = \gamma = \delta = e = 1$, $c = 3$, and $m = 0.2$, $m = 0.1$, correspondingly

Theorem 1

1. *Equilibrium B is unstable;*
2. *Equilibrium A is stable for* $(c, m) \in DA$ *if* $0 < c < c_h$*, is unstable for* $(c, m) \in DA$ *if* $c_h < c < c_0$ *and for* $(c, m) \in DC_1$*, it leaves nonnegative octant if* $c > c_0$*;*
3. *Equilibrium* C^+ *is positive and stable only for* $(c, m) \in DC$*;*
4. *Equilibrium* C^- *is positive and unstable for* $(c, m) \in DC_2$*,* C^- *is negative or does not exist for* $(c, m) \notin DC_2$*;*
5. *Equilibrium O is non-hyperbolic, projection of its neighborhood to the plane* $p = 0$ *contains only hyperbolic sector for* $0 < c < c_b$*, contains hyperbolic and stable parabolic sectors for* $c_b < c < c_0$*, contains elliptic sector for* $c > c_0$ [1].

Based on Theorem 1 and computer analysis we present the bifurcation diagram of system (1) in Fig. 2. It shows only stable modes in the (N, z)-phase portraits of the model.

Theorem 2

1. *The parameter domain DA in (c, m)-plane is dividing into three subdomains 1A, 2A, 3A. The equilibrium A is globally stable in 1A, does not exist in Domain 3A where O is a single equilibrium and has elliptic sector in its positive neighborhood; Domain 2A is a domain of bistability, where either equilibrium A or a limit cycle is stable and share basins with equilibrium O.*
2. *The parameter domain DC in (c, m)-plane is dividing into three subdomains 1C, 2C and 3C of different $(3D)$-phase portraits. The equilibrium C is globally stable in 1C; it shares basins with O in 2C. In Domain 3C (N, z)-phase portrait of system (1) contains the stable equilibrium C^+ and equilibrium O,which has an attractive sector and/or elliptic sector in its neighborhood.*

4 Discussion and Conclusion

Our analysis of the "consumer–predator-renewable resource" model 1 shows that predators can essentially change the dynamics and steady states of "predator-free" system. Predators do not change essentially the dynamics of the consumer-renewable resource system when the level of over-consumption is not too large (see Domains 1, 2 in Fig. 2). In contrast, predators are able to keep a stable equilibrium with nonzero amounts of preys and resource even when the level of over-consumption is large so that the "predator-free" consumer–resource system get extinction (Domain 3C in Fig. 2). Note that the amount of predators in this equilibrium increases as the parameter c increases. The equilibrium point has a bounded basin and trajectories starting out this basin tend to O, i.e., the system goes to extinct.

Computer experiments revealed that even small amount of predators in the system increases the duration of the system "existence" even in the case when in the "final" equilibrium the amount of predators is zero.

Overall, we may conclude that the model reveals possible "positive" influence of predators which can increase sustainability of the system and prevent it from extinction. It is our hope that the model and the results of its studying may be interpreted in terms of social–economical systems, but it is out of the scope of this work.

References

1. Berezovskaya, F.S., Novozhilov, A.S., Karev, G.P.: Population models with singular equilibrium. Math. Biosci. **208**(1), 270–299 (2007)
2. Hardin, G.: The tragedy of the commons: Science **162**(5364), 1243–1248 (1968)
3. Kareva, I., Berezovskaya, F., Castillo-Chavez, C.. Transitional regimes as early warning signals in resource dependent competition models, Math. Biosci. **240,** 114–123 (2012)

4. Krakauer, D.C., Page, K.M., Erwin, D.H.: Diversity, dilemmas, and monopolies of niche construction. Am. Nat. **173**(1), 26–40 (2009)
5. Odling-Smee, F.J., Laland, K.N., Feldman, M.W.: Niche Construction: The Neglected Process in Evolution (MPB-37). Princeton University Press (2003)
6. Volterra, V.: Leçons sur la Théorie Mathématique de la Lutte pour la Vie. Gauthier-Villare, Paris (1931).

Strip Saturation Yield Model for a Piezoelectric Plate: A Study on Influence of Change in Poling Direction

R.R. Bhargava and Kamlesh Jangid

Abstract A study on the influence of change in poling direction is carried for mechanical and electric strip yield model for a transversely isotropic piezoelectric plate cut along two equal collinear semipermeable cracks. Solution is obtained using Stroh formalism and complex variable technique. An illustrative numerical example is considered for a poled PZT-5H ceramic plate to show the effect of change in poling direction on energy release rate (ERR).

1 Introduction

A vast variety of crack problems for piezoelectric ceramics have been investigated considering impermeable, permeable electric conditions on the crack faces. It is noted that these conditions give higher and lower estimate of energy release rate (ERR), respectively. But empirically it is observed that semipermeable crack face boundary condition give more accurate results. Semipermeable boundary conditions may be defined as (given by Hao and Shen [5])

$$D_2^+ = D_2^-, \qquad D_2^+(u_2^+ - u_2^-) = \varepsilon_a(\phi^- - \phi^+), \tag{1}$$

where D_2, u_2, ϕ, and ε_a, respectively, denote the electric-displacement, mechanical displacement components perpendicular to the crack, electric potential, and permittivity of media inside crack gap. The superscripts $+$ and $-$ denote the value of quantity over upper and lower faces of the cracks.

A simple model for a slit arrest under small scale mechanical yielding proposed by Dugdale [3] was also extended for both mechanical and electric saturation model for piezoceramics. Shen et al. [8] obtained solution for a strip electric saturation

K. Jangid (✉)
Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: kamljdma@iitr.ac.in

R.R. Bhargava
Department of Mathematics, Indian Institute of Technology Roorkee,
Roorkee 247667, India
e-mail: rajrbfma@iitr.ac.in

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_12

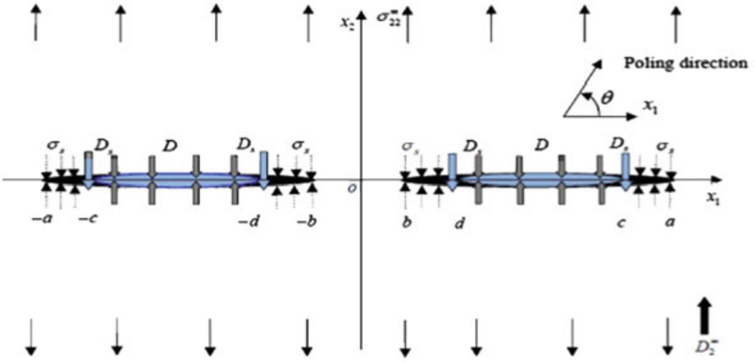


Fig. 1 Schematic configuration of the problem

and mechanical yielding model for an interface crack between ferroelectric–plastic bimaterials. A mechanical and electric yield model for impermeable crack in a piezoelectric ceramic had been investigated by Loboda et al. [6].

More recently, we [1] have given the solution of strip-electromechanical model for piezoelectric plate cut along two semipermeable collinear cracks. The influence of poling direction and electric boundary conditions on fracture behavior of a finite crack in two-dimensional infinite piezoelectric medium had been studied by Fan and Zhao [4].

Very few studies are available for changing poling direction. Therefore, we address this paucity, studying the influence of change in poling direction on a piezoelectric media cut along two equal collinear straight cracks under in-plane electrical and mechanical loads.

2 Statement and Solution of the Problem

A transversely isotropic poled piezoelectric plate weakened by two equal collinear semipermeable cracks, which occupy the region $x_2 = 0, d \leq |x| \leq c$ in $x_1 o x_2$ plane. The poling direction makes an angle θ with x_1-axis. The remote boundary of the plate is subjected to in-plane normal, uniform constant tension $\sigma_{22} = \sigma_{22}^{\infty}$, and electrical displacement $D_2 = D_2^{\infty}$, consequently cracks open in self-similar fashion forming a strip yield and saturation zone of equal length ahead each tip of the crack. The developed zones at the cracks tips $c, d, -d$, and $-c$, occupy the respective intervals $[c, a], [b, d], [-d, -b]$, and $[-a, -c]$ on ox_1-axis. To stop the crack from further opening, the rims of the developed zones are subjected to uniform constant normal cohesive yield point stress $\sigma_{22} = \sigma_s$ and saturation limit in-plane electric displacement $D_2 = D_s$. The schematic representation of the problem is depicted in Fig. 1.

The physical boundary conditions stated above may mathematically be written as

(i) $\sigma_{22}^{+} = \sigma_{22}^{-} = 0,\, D_2 = D,$ on $L = \cup_1^2 L_i$
(ii) $\sigma_{22} = \sigma_{22}^{\infty},\, D_2 = D_2^{\infty},$ for $|x_2| \to \infty$
(iii) $\sigma_{22}^{+} = \sigma_{22}^{-} = \sigma_s,\, D_2^{+} = D_2^{-} = D_s,$ for $b \leq |x_1| \leq d, c \leq |x_1| \leq a$
(iv) $\Phi_{,1}^{+}(x_1) = \Phi_{,1}^{-}(x_1) = -\mathbf{V}, \mathbf{V} = [0, \sigma_{22}^{\infty}, 0, D_2^{\infty}]^T$ on $d < |x_1| < c,$

where D is the electric flux through the crack regions $[-c, -d]$ and $[d, c]$, determined using Eq. (1).

According to Stroh formalism, the solution of the problem may be given as

$$\mathbf{u}_{,1} = \mathbf{AF}(z) + \overline{\mathbf{AF}(z)}, \tag{2}$$

$$\Phi_{,1} = \mathbf{BF}(z) + \overline{\mathbf{BF}(z)}. \tag{3}$$

The methodology presented here is recapitulated from Bhargava and Kamlesh [2] to make the chapter self-sufficient for a reader.

The continuity of $\Phi_{,1}(x_1)$ on the whole real axis implies that

$$[\mathbf{BF}(x_1) - \overline{\mathbf{BF}}(x_1)]^{+} - [\mathbf{BF}(x_1) - \overline{\mathbf{BF}}(x_1)]^{-} = \mathbf{0}. \tag{4}$$

According to Muskhelishvili [7], its solution may be written as

$$\mathbf{BF}(z) = \overline{\mathbf{BF}}(z) = \mathbf{h}(z)(\text{say}). \tag{5}$$

Boundary condition (iv) together with Eqs. (3) and (5) yield following Hilbert problem

$$\mathbf{h}^{+}(x_1) + \mathbf{h}^{-}(x_1) = \mathbf{V}^0 - \mathbf{V}, \qquad \mathbf{V}^0 = [0, 0, 0, D]^T, \qquad d < |x_1| < c. \tag{6}$$

Introducing a new complex function vector $\Omega(z) = [\Omega_1, \Omega_2, \Omega_3, \Omega_4]^T$ as $\Omega(z) = \mathbf{H}^R\mathbf{BF}(z)$, which, using Eq. (5), gives the relation $\mathbf{h}(z) = \Lambda\Omega(z)$, where $\Lambda = [\mathbf{H}^R]^{-1}$, $\mathbf{H}^R = 2Re\mathbf{Y}, \mathbf{Y} = i\mathbf{AB}^{-1}$.

Consequently, Eq. (6) may be written in component form for $\Omega_2(z)$ and $\Omega_4(z)$, as

$$\Lambda_{22}[\Omega_2^{+}(x_1) + \Omega_2^{-}(x_1)] + \Lambda_{24}[\Omega_4^{+}(x_1) + \Omega_4^{-}(x_1)] = -\sigma_{22}^{\infty}, \quad d < |x_1| < c, \tag{7}$$

$$\Lambda_{42}[\Omega_2^{+}(x_1) + \Omega_2^{-}(x_1)] + \Lambda_{44}[\Omega_4^{+}(x_1) + \Omega_4^{-}(x_1)] = D - D_2^{\infty}, \quad d < |x_1| < c. \tag{8}$$

The solution of above Hilbert problems may be written, using Muskhelishvili [7], as

$$\Omega_2(z) = \frac{\Lambda_{44}\sigma_{22}^{\infty} + \Lambda_{24}(D - D_2^{\infty})}{2\Delta}\left\{\frac{z^2 - a^2\lambda_2^2}{X_2(z)} - 1\right\} - \frac{\Lambda_{44}\sigma_s + \Lambda_{24}(D - D_s)}{\pi\Delta X_2(z)}R, \tag{9}$$

$$\Omega_4(z) = \frac{\Lambda_{42}\sigma_{22}^{\infty} + \Lambda_{22}(D - D_2^{\infty})}{2\Delta}\left\{1 - \frac{z^2 - a^2\lambda_2^2}{X_2(z)}\right\} + \frac{\Lambda_{42}\sigma_s + \Lambda_{22}(D - D_s)}{\pi\Delta X_2(z)}R, \tag{10}$$

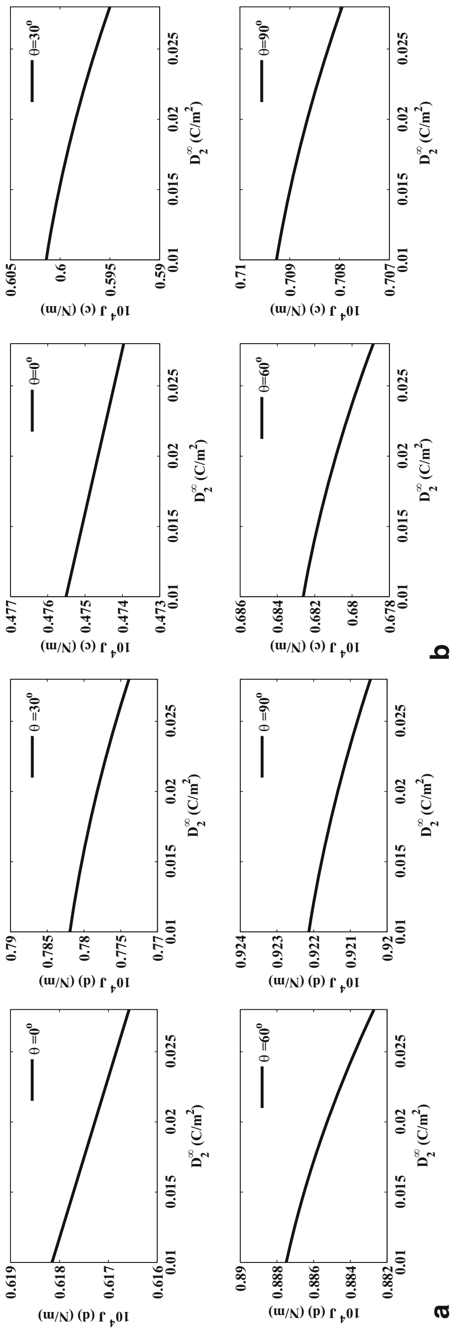


Fig. 2 Energy release rate (ERR) versus prescribed D_2^∞

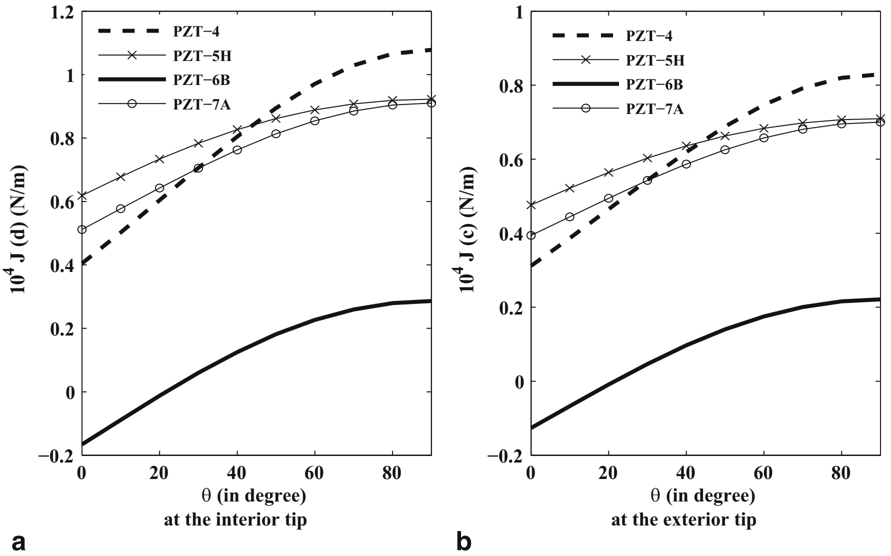


Fig. 3 Energy release rate (ERR) versus poling angle

where $X_2(z) = \sqrt{(z^2-a^2)(z^2-b^2)}$, $\Delta = \Lambda_{22}\Lambda_{44} - \Lambda_{24}\Lambda_{42}$,
$R = \left\{(z^2 - a^2\lambda_2^2)(\frac{\pi}{2} - \vartheta_d + \vartheta_c) - X_2(z)(\frac{\pi}{2} - \upsilon_d + \upsilon_c) + R_1\right\}$,
$R_1 = da\left(E(\vartheta_d, k_2) - \lambda_2^2 F(\vartheta_d, k_2)\right) - ca\left(E(\vartheta_c, k_2) - \lambda_2^2 F(\vartheta_c, k_2)\right) - (a^2 - b^2)$
$(\sin\vartheta_d \cos\vartheta_d - \sin\vartheta_c \cos\vartheta_c)$,
$k_2^2 = 1 - (b/a)^2$, $\lambda_2^2 = E(k_2)/F(k_2)$, $\sin^2\vartheta_d = (a^2 - d^2)/(a^2 - b^2)$,
$\sin^2\vartheta_c = (a^2 - c^2)/(a^2 - b^2)$,
$\upsilon_d = \tan^{-1}\sqrt{\frac{(b^2-z^2)(a^2-d^2)}{(a^2-z^2)(d^2-b^2)}}$, $\upsilon_c = \tan^{-1}\sqrt{\frac{(b^2-z^2)(a^2-c^2)}{(a^2-z^2)(c^2-b^2)}}$.

3 Applications

The size of developed zones is obtained under the Dugdale hypothesis of stresses, $\sigma_{22}(x_1)$, and electric displacement, $D_2(x_1)$, remain finite at the tips $x_1 = b$ and $x_1 = a$, then one obtains nonlinear equations to determine b and a from

$$\left(\frac{b^2}{a^2} - \lambda_2^2\right)\left(\frac{\pi}{2}L - \vartheta_{dc}\right) - \frac{R_1}{a^2} = 0, \tag{11}$$

$$(1 - \lambda_2^2)\left(\frac{\pi}{2}L - \vartheta_{dc}\right) - \frac{R_1}{a^2} = 0, \tag{12}$$

where $L = \sigma_{22}^{\infty}/\sigma_s$ or $(D - D_2^{\infty})/(D - D_s)$, and $\vartheta_{dc} = \frac{\pi}{2} - \vartheta_d + \vartheta_c$.

The relative opening of the crack faces, $\triangle u_2$ at the tips d and c may be given as

$$\triangle u_2(d) = -\frac{\varLambda_{44}\sigma_s + \varLambda_{24}(D - D_s)}{\pi\varDelta}\left\{R_2 - \frac{2R_1}{a^2}F(\xi_d, k_2) + 2a\vartheta_{dc}R_3 + R_4\right\} + a\frac{\varLambda_{44}\sigma_{22}^{\infty} + \varLambda_{24}(D - D_2^{\infty})}{\varDelta}\big(R_3 - \lambda_2^2 F(\xi_d, k_2)\big), \tag{13}$$

$$\triangle u_2(c) = \frac{\varLambda_{44}\sigma_s + \varLambda_{24}(D - D_s)}{\pi\varDelta}\left\{R_5 + R_6 - \frac{2R_1}{a^2}F(\vartheta_c, k_2) + 2a\vartheta_{dc}E(\vartheta_c, k_2)\right\} - a\frac{\varLambda_{44}\sigma_{22}^{\infty} + \varLambda_{24}(D - D_2^{\infty})}{\varDelta}\big(E(\vartheta_c, k_2) - \lambda_2^2 F(\vartheta_c, k_2)\big). \tag{14}$$

Also, the jump in electric potential across the two faces of the cracks at tips d and c may be given as

$$\triangle u_4(d) = \frac{\varLambda_{42}\sigma_s + \varLambda_{22}(D - D_s)}{\pi\varDelta}\left\{R_2 - \frac{2R_1}{a^2}F(\xi_d, k_2) + 2a\vartheta_{dc}R_3 + R_4\right\} - a\frac{\varLambda_{42}\sigma_{22}^{\infty} + \varLambda_{22}(D - D_2^{\infty})}{\varDelta}\big(R_3 - \lambda_2^2 F(\xi_d, k_2)\big), \tag{15}$$

$$\triangle u_4(c) = -\frac{\varLambda_{42}\sigma_s + \varLambda_{22}(D - D_s)}{\pi\varDelta}\left\{R_5 + R_6 - \frac{2R_1}{a^2}F(\vartheta_c, k_2) + 2a\vartheta_{dc}E(\vartheta_c, k_2)\right\} + a\frac{\varLambda_{42}\sigma_{22}^{\infty} + \varLambda_{22}(D - D_2^{\infty})}{\varDelta}\big(E(\vartheta_c, k_2) - \lambda_2^2 F(\vartheta_c, k_2)\big), \tag{16}$$

where $R_2 = -d\ln\left(\frac{a^2-d^2}{a^2-b^2} + \frac{a^2(b^2-d^2)}{d^2(a^2-b^2)}\right) + \frac{2b^2}{a}\sqrt{\frac{a^2-d^2}{d^2-b^2}}\left(F(\xi_d, k_2) - \varPi(\xi_d, \frac{d^2-b^2}{d^2}, k_2)\right)$,
$R_3 = E(\xi_d, k_2) - \frac{k_2^2\sin\xi_d\cos\xi_d}{\sqrt{1-k_2^2\sin^2\xi_d}}, \quad \sin^2\xi_d = \frac{a^2(d^2-b^2)}{d^2(a^2-b^2)}$,
$R_4 = d\ln\left(\frac{\sqrt{(d^2-b^2)(a^2-c^2)}+\sqrt{(a^2-d^2)(c^2-b^2)}}{\sqrt{(d^2-b^2)(a^2-c^2)}-\sqrt{(a^2-d^2)(c^2-b^2)}}\right) - \frac{2b^2}{a}\sqrt{\frac{a^2-c^2}{c^2-b^2}}\varPi(\xi_d, \frac{c^2k_2^2}{c^2-b^2}, k_2)$,
$R_5 = -c\ln\left(\frac{(a^2-c^2)(c^2-b^2)}{c^2(a^2-b^2)} + 1\right) + \frac{2}{a}\sqrt{\frac{c^2-b^2}{a^2-c^2}}\left(F(\vartheta_c, k_2) - c^2\varPi(\vartheta_c, \frac{a^2-c^2}{a^2}, k_2)\right)$,
$R_6 = -\frac{2}{a}\sqrt{(d^2-b^2)(a^2-d^2)}\left(F(\vartheta_c, k_2) + \frac{d^2}{a^2-d^2}\varPi(\vartheta_c, \frac{a^2-b^2}{a^2-d^2}, k_2)\right)$
$+ c\ln\left(\frac{\sqrt{(c^2-b^2)(a^2-d^2)}+\sqrt{(a^2-c^2)(d^2-b^2)}}{\sqrt{(c^2-b^2)(a^2-d^2)}-\sqrt{(a^2-c^2)(d^2-b^2)}}\right)$.

To find the value of electric flux D, the quadratic equation is obtained from Eq. (1) by substituting $\triangle u_2(x_1)$ and $\triangle u_4(x_1)$ for two collinear equal cracks problem, as

$$\eta_1 D^2 + \eta_2 D + \eta_3 = 0, \tag{17}$$

where $\eta_1 = \varLambda_{24}$, $\eta_2 = \varLambda_{44}\sigma_{22}^{\infty} - D_2^{\infty}\varLambda_{24} - \varepsilon_a\varLambda_{22}$, $\eta_3 = -\varepsilon_a(\varLambda_{42}\sigma_{22}^{\infty} - D_2^{\infty}\varLambda_{22})$.

The value of D is chosen for which $\triangle u_2(x_1)$ is positive, and the ERR at the interior and exterior tips of the crack is calculated using

$$J(d) = \sigma_s\triangle u_2(d) + D_s\triangle u_4(d), \qquad J(c) = \sigma_s\triangle u_2(c) + D_s\triangle u_4(c). \tag{18}$$

4 Results and Discussions

Figure 2a depicts the variation of ERR versus prescribed electrical load, D_2^{∞}, for PZT-5H ceramic at the interior tip d. It may be noted that ERR decreases even as D_2^{∞} is increased for all poling direction. However, the ERR is minimum when poling is along the length of the crack, while it is maximum when poling direction is perpendicular to crack length. A similar variation is plotted in Fig. 2b at the exterior tip c. It is important to note the ERR (at exterior) is less vis-a-vis that at interior.

Figure 3a and b show the ERR variation with respect to poling direction angle for different poled piezoceramics at the interior and exterior crack tips. This variation is useful for the selection of desired ceramic for the specific work.

5 Conclusions

It is seen that poling direction has a definite effect on ERR, consequently on crack opening arrest. ERR is minimum when poling direction is along crack length and as it is changed to 90° to crack length, the ERR is increased.

References

1. Bhargava, R.R., Jangid, K.: Strip-electromechanical model solution for piezoelectric plate cut along two semi-permeable collinear cracks. Arch. Appl. Mech. **83**, 1469–1491 (2013)
2. Bhargava, R.R., Jangid, K.: A study on influence of poling direction on piezoelectric plate weakened by two collinear semi-permeable cracks. Acta Mech. **225**, 109–129 (2014)
3. Dugdale, D.S.: Yielding of steel sheets containing slits. J. Mech. Phys. Solids **8,** 100–104 (1960)
4. Fan, C.Y., Zhao, M.H.: Influence of poling direction and boundary conditions on cracks in 2D piezoelectric media. IEEE conference (2010)
5. Hao, T.H., Shen, Z.Y.: A new electric boundary condition of electric fracture mechanics and its applications. Eng. Fract. Mech. **47,** 793–802 (1994)
6. Loboda, V., Lapusta, Y., Govorukha, V.: Mechanical and electrical yielding for an electrically insulated crack in an interlayer between piezoelectric materials. Int. J. Eng. Sci. **46,** 260–272 (2008)
7. Muskhelishvili, N.I.: Some Basic problems of Mathematical Theory of Elasticity. Noordhoff Leyden (1975)
8. Shen, S., Nishioka, T., Kuang, Z.B., Liu, Z.: Nonlinear electromechanical interfacial fracture for piezoelectric materials. Mech. Mater. **32,** 57–64 (2000)

Strip-Saturation-Induction Model Mode-III Solution for Piezoelectromagnetic Strip

R. R. Bhargava and Pooja Raj Verma

Abstract Using Fourier cosine integral transform technique, mode-III strip-induction-saturation model is proposed for a cracked transversely isotropic piezoelectromagnetic strip. Strip edges are subjected to combined anti-plane mechanical and in-plane electromagnetic loadings. Analytical closed-form expressions are derived for developed zones, field intensity factors, and energy release rates. Four impermeable/permeable electromagnetic crack-faces boundary conditions are considered. Results are plotted for $BaTiO_3$–$CoFe_2O_4$ ceramic to show the influence of electromagnetic fields on local energy release rate (LERR) and global energy release rate(GERR).

1 Introduction

Due to intrinsic coupling among elastic, electric, and magnetic field, magnetoelectroelastic (MEE) ceramic is widely used in medical and industrial engineering as sensor, actuators or transducers, etc.

The fracture behavior of MEE depends on applied loading as well as electromagnetic crack-faces boundary conditions. Electromagnetic crack-face boundary conditions are the most important and basic issues in studying the facture behavior of MEE materials. Electromagnetic crack-face boundary conditions for MEE ceramic are given by Wang and Mai [1].

A strip-electric saturation model proposed by Gao et al. [2] for a piezoelectric ceramic. And this model extended for both electrically and mechanically yield model by Shen et al. [3]. More recently, Ma et al. [4] proposed a contact zone model for two dissimilar MEE materials, with electrically impermeable (EI) and magnetically permeable (MP) crack-face boundary conditions.

P. R. Verma (✉) · R. R. Bhargava
Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: prajvdma@iitr.ac.in, poojarajvs@gmail.com

R. R. Bhargava
e-mail: rajrbfma@iitr.ac.in

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_13

In this chapter, we have obtained mode-III strip-induction-saturation mathematical model solution for cracked piezoelectromagnetic strip. The crack-faces are assumed to be mechanically traction free. Also, the inside crack gap media is assumed to be (a) electromagnetically impermeable, (b) electrically permeable (EP) and magnetically impermeable (MI), (c) EI and MP, and (d) electromagnetically permeable.

2 Statement and Solution of the Problem

Consider a MEE strip occupies the region $-h_2 \leq y \leq h_1$ and $|x| < \infty$ in xoy-plane and is thick enough in z-direction to allow anti-plane deformation. The strip is both electrically and magnetically poled along z-direction. The MEE strip is weakened by a non-centric crack which occupies the region $y = 0, -a \leq x \leq a$ and oriented parallel to the strip edges. Uniform constant anti-plane mechanical load, $\sigma_{zy} = \tau_0$, in-plane electrical, $D_y = D_0$, and magnetic loads, $B_y = B_0$, are prescribed on edges of the strip, opens the crack in self-similar fashion. Consequently, under small-scale-electromagnetic yielding, saturation and induction zones develop ahead of each crack tips (which are assumed to be of equal length), occupy the interval $c \leq |x| \leq a$ on ox-axis. Rims of the developed zones are subjected to in-plane normal cohesive saturation limit electric-displacement $D_y = D_s$ and induction limit magnetic induction $B_y = B_s$, to stop the crack from further opening.

Under mode-III deformation, the *constitutive equations* for MEE media may be written as
$\sigma_{zy} = c_{44}\gamma_{zy} + e_{15}\phi_{,y} + h_{15}\psi_{,y}, D_y = e_{15}\gamma_{zy} - \varepsilon_{11}\phi_{,y} - d_{11}\psi_{,y},$
$B_y = h_{15}\gamma_{zy} - d_{11}\phi_{,y} - \mu_{11}\psi_{,y},$
where ϕ, ψ, γ_{zy}, $c_{44}, e_{15}, \varepsilon_{11}, h_{15}$, and μ_{11} are electric potential, magnetic potential, strain, elastic constant, piezoelectric constant, dielectric constant, piezomagnetic coefficient, and electromagnetic coefficient, respectively.

The MEE boundary conditions combined over the extended crack surfaces using superposition principle and continuity conditions along the line $y = 0$ may be written as

$$\begin{array}{llll}
\text{(i)} & \tau_{zy(1)}(x,0^+) = \tau_{zy(2)}(x,0^-) = -\tau_0, & & 0 \leq x \leq a, \\
\text{(ii)} & D_{y(1)}(x,0^+) = D_{y(2)}(x,0^-) = -D_0(1-D) + D_s H(x-a), & & 0 \leq x \leq c, \\
\text{(iii)} & B_{y(1)}(x,0^+) = B_{y(2)}(x,0^-) = -B_0(1-B) + B_s H(x-a), & & 0 \leq x \leq c, \\
\text{(iv)} & w_{(1)}(x,0^+) = w_{(2)}(x,0^-), & \tau_{zy(1)}(x,0^+) = \tau_{zy(2)}(x,0^-), & |x| > a, \\
\text{(v)} & \phi_{(1)}(x,0^+) = \phi_{(2)}(x,0^-), & D_{y(1)}(x,0^+) = D_{y(2)}(x,0^-), & |x| > c, \\
\text{(vi)} & \psi_{(1)}(x,0^+) = \psi_{(2)}(x,0^-), & B_{y(1)}(x,0^+) = B_{y(2)}(x,0^-), & |x| > c,
\end{array}$$

where $H(.)$ denotes Heaviside function. D and B are the electric and magnetic flux inside the crack $[-a, a]$ with zero value for impermeable crack and approximately one for permeable crack. Subscripts (1) and (2) refer to MEE material layers for upper $0 < y \leq h_1$ and lower $-h_2 \leq y < 0$ regions of the strip, respectively. Because of the symmetry in geometry and loading, it is sufficient to consider the problem for $0 \leq x < \infty$ region only.

For convenience of mathematics, we introduce two new potential functions $\Phi(x, y)$ and $\Psi(x, y)$ as

$$\phi(x, y) = \Phi(x, y) + m_1\Psi(x, y) + m_2 w(x, y), \tag{1}$$

$$\psi(x, y) = \Psi(x, y) + m_3\Phi(x, y) + m_4 w(x, y), \tag{2}$$

where $m_1 = -\frac{d_{11}}{\varepsilon_{11}}$, $m_2 = \frac{d_{11}h_{15}-e_{15}\mu_{11}}{d_{11}^2-\mu_{11}\varepsilon_{11}}$, $m_3 = -\frac{d_{11}}{\mu_{11}}$, and $m_4 = \frac{d_{11}e_{15}-\varepsilon_{11}h_{15}}{d_{11}^2-\mu_{11}\varepsilon_{11}}$. The governing equations for the problem may be written as

$$\nabla^2 w(x, y) = 0, \nabla^2\Phi(x, y) = 0, \nabla^2\Psi(x, y) = 0,$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is a two-dimensional Laplacian operator.

The solution of Laplace equations may be written using Fourier integral transform technique as

$$w_{(i)}(x, y) = \int_0^\infty [A_{(i)}(\xi)\cosh(y\xi) + B_{(i)}(\xi)\sinh(y\xi)]\cos(x\xi)d\xi, \tag{3}$$

$$\Phi_{(i)}(x, y) = \int_0^\infty [C_{(i)}(\xi)\cosh(y\xi) + E_{(i)}(\xi)\sinh(y\xi)]\cos(x\xi)d\xi, \tag{4}$$

$$\Psi_{(i)}(x, y) = \int_0^\infty [F_{(i)}(\xi)\cosh(y\xi) + G_{(i)}(\xi)\sinh(y\xi)]\cos(x\xi)d\xi, \tag{5}$$

Where $i = 1, 2$ represents the upper and lower region of the MEE strip, respectively.

Using boundary conditions (i–vi) into constitutive Eqs. (1) and (2), we obtain three set of dual integral equations that can be further reduced into three Fredholm integral equation of the seconds kind

$$\Omega_1(x) + \int_0^1 K_1(x, t)\Omega_1(t)dt = \sqrt{x}\frac{\overline{\varepsilon}_{11}\overline{\mu}_{11}\tau_0 + \overline{\mu}_{11}\overline{e}_{15}D_0(1-D) + \overline{\varepsilon}_{11}\overline{h}_{15}B_0(1-B)}{\overline{c}_{44}\overline{\mu}_{11}\overline{\varepsilon}_{11}}, \tag{6}$$

$$\Omega_2(x) + \int_0^1 K_2(x, t)\Omega_2(t)dt = \frac{\sqrt{x}}{\overline{\varepsilon}_{11}}D_0(1-D) - \frac{2\sqrt{x}}{\pi\overline{\varepsilon}_{11}}D_s\cos^{-1}\left(\frac{a}{cx}\right), 0 \le x \le c, \tag{7}$$

$$\Omega_3(x) + \int_0^1 K_3(x, t)\Omega_3(t)dt = \frac{\sqrt{x}}{\overline{\mu}_{11}}B_0(1-B) - \frac{2\sqrt{x}}{\pi\overline{\mu}_{11}}B_s\cos^{-1}\left(\frac{a}{cx}\right), 0 \le x \le c, \tag{8}$$

by introducing $A_{(1)}(\xi)$, $C_{(1)}(\xi)$, and $F_{(1)}(\xi)$ in term of new auxiliary function $\Omega_1(.)$, $\Omega_2(.)$, and $\Omega_3(.)$, respectively, as

$$A_{(1)}(\xi) = \frac{2a^2}{[1 + M_{12}(\xi)]}\int_0^1 \sqrt{u}\Omega_1(u)J_0(au\xi)du, \tag{9}$$

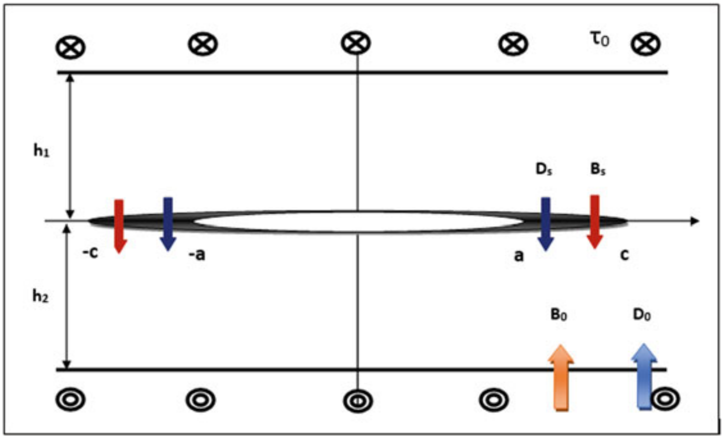


Fig. 1 Schematic representation of the problem

$$C_{(1)}(\xi) = -\frac{2c^2}{[1+M_{12}(\xi)]}\int_0^1 \sqrt{u}\Omega_2(u)J_0(cu\xi)du, \tag{10}$$

$$F_{(1)}(\xi) = -\frac{2c^2}{[1+M_{12}(\xi)]}\int_0^1 \sqrt{u}\Omega_3(u)J_0(cu\xi)du, \tag{11}$$

where $J_0(.)$ be modified zero-order Bessel function of the first kind. The kernels $K_1(x,t)$, $K_2(x,t)$, and $K_3(x,t)$ are given as $K_1(x,t) = \sqrt{xt}\int_0^\infty y[\eta(y/a) - 1]J_0(xy)J_0(ty)dy$, $K_2(x,t) = K_3(x,t) = \sqrt{xt}\int_0^\infty y[\eta(y/c) - 1]J_0(xy)J_0(ty)dy$, and $\eta(y) = \frac{2\tanh(h_1\xi)}{1+\tanh(h_1\xi)\coth(h_2\xi)}$.

3 Applications and Case Study

Stress intensity factor, K_{III}^{τ}, strain intensity factor, K_{III}^{γ}, electric displacement intensity factor, K_{IV}^{D}, electric field intensity factor, K_{IV}^{E}, magnetic induction intensity factor, K_{V}^{B}, and magnetic field intensity factor, K_{V}^{H}, are obtained as

$$K_{III}^{\tau} = \lim_{x\to a^+}\sqrt{2\pi(x-a)}\tau_{zy}(x,0) = \overline{c}_{44}\sqrt{\pi a}\Omega_1(1), \tag{12}$$

$$K_{III}^{\gamma} = \lim_{x\to a^+}\sqrt{2\pi(x-a)}\gamma_{zy}(x,0) = \sqrt{\pi a}\Omega_1(1), \tag{13}$$

$$K_{IV}^{D} = \lim_{x\to c^+}\sqrt{2\pi(x-c)}D_y(x,0) = \overline{\varepsilon}_{11}\sqrt{\pi c}\Omega_2(1), \tag{14}$$

$$K_{IV}^{E} = \lim_{x\to c^+}\sqrt{2\pi(x-c)}E_y(x,0) = \sqrt{\pi c}\Omega_2(1), \tag{15}$$

$$K_{V}^{B} = \lim_{x\to c^+}\sqrt{2\pi(x-c)}B_y(x,0) = \overline{\mu}_{11}\sqrt{\pi c}\Omega_3(1), \tag{16}$$

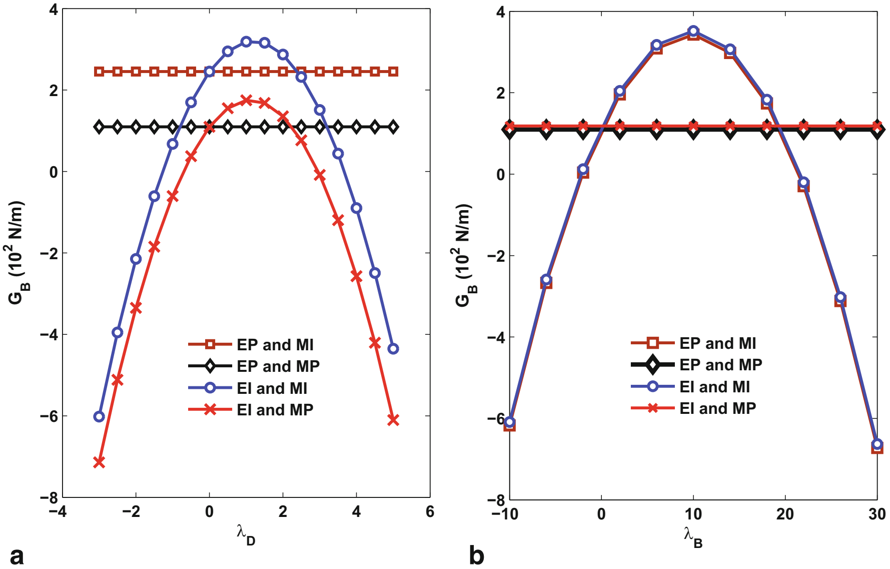


Fig. 2 Variation in global energy release rate (GERR) with respect to **a** λ_D and **b** λ_B

$$K_V^H = \lim_{x \to c^+} \sqrt{2\pi(x-c)}H_y(x,0) = \sqrt{\pi c}\Omega_3(1). \tag{17}$$

The following relation gives the condition of magnetic and electric forces required to produce equal electric and magnetic yielding

$$B_0 D_s(1-B) - D_0 B_s(1-D) - \overline{\mu}_{11} R_1 D_s + \overline{\varepsilon}_{11} R_2 B_s = 0, \tag{18}$$

where $R_1 = \int_0^1 K_3(1,t)\Omega_3(t)dt$ and $R_2 = \int_0^1 K_2(1,t)\Omega_2(t)dt$.
$\overline{c}_{44} = c_{44} + \frac{2h_{15}d_{11}e_{15} - \varepsilon_{11}h_{15}^2 - \mu_{11}e_{15}^2}{d_{11}^2 - \mu_{11}\varepsilon_{11}}$, $\overline{e}_{15} = e_{15} - \frac{d_{11}h_{15}}{\mu 11}$, and $\overline{\varepsilon}_{11} = \varepsilon_{11} - \frac{d_{11}^2}{\mu_{11}}$.

Local, G_L, and global, G_B, energy release rates at the crack tip, $x = a$, may be respectively, given as

$$G_L = (K_{III}^{\tau} K_{III}^{\gamma})/2 = [\overline{c}_{44}(\pi a)\{\Omega_1(1)\}^2]/2, \tag{19}$$

$$\begin{aligned} G_B &= \{K_{III}^{\tau} K_{III}^{\gamma} - K_{IV}^{D} K_{IV}^{E} - K_V^B K_V^H\}/2 \\ &= [\overline{c}_{44}(\pi a)\{\Omega_1(1)\}^2 - \overline{\varepsilon}_{11}(\pi c)\{\Omega_2(1)\}^2 - \overline{\mu}_{11}(\pi c)\{\Omega_3(1)\}^2]/2. \end{aligned} \tag{20}$$

A numerical case study is presented for a transversely isotropic $BaTiO_3$–$CoFe_2O_3$ material to investigate the effects of different types of crack-face boundary conditions on local energy release rate (LERR) and global energy release rate (GERR). The prescribed mechanical load and crack length are assumed to be 100 MPa and 20 mm, respectively.

The material constants for $BaTiO_3$–$CoFe_2O_3$ are given as $c_{44} = 44$ GPa, $e_{15} = 5.8$ C/m^2, $h_{15} = 275$ N/Am, $\varepsilon_{11} = 5.367 \times 10^{-9}$ Ns/VC, $\mu_{11} = \times 10^{-4}$ Ns/Vm, and $d_{11} = 2.97 \times 10^{-12}$ Ns^{2}C^2.

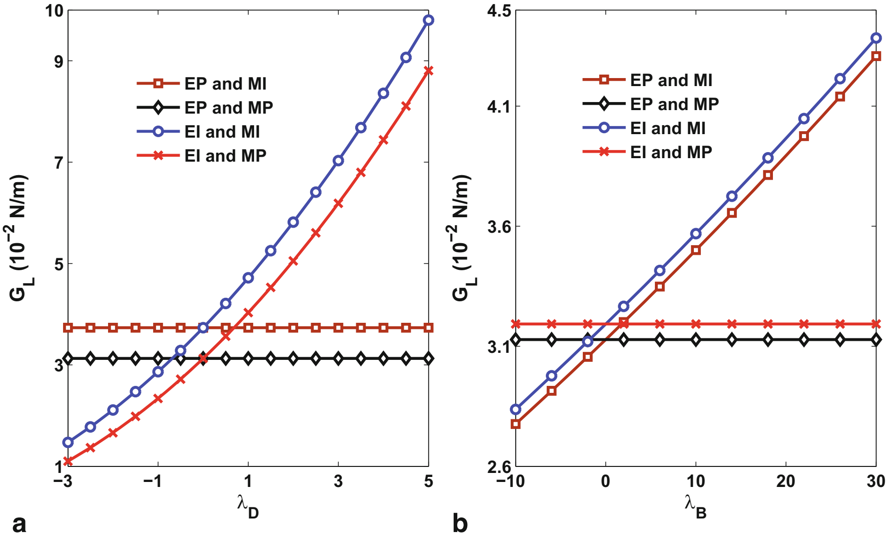


Fig. 3 Variation in local energy release rate (LERR) with respect to **a** λ_D and **b** λ_B

Figures 2(a) and 2(b), respectively, depicts the behavior of GERR with respect to the variation in the nondimensional electrical, $\lambda_D = c_{44}D_0/e_{15}\tau_0$, and magnetic loading coefficient $\lambda_B = c_{44}B_0/h_{15}\tau_0$ for different electromagnetic crack-face boundary conditions. In Fig. 2(a) for case (EI and MI) and (EI and MP), it is pointed that GERR decreases symmetrically for $\lambda_D \approx 1$. And in Fig. 2(b), GERR is symmetrical about $\lambda_B \approx 10$ for case (EP and MI) and (EP and MI). This shows that GERR is independent of the direction of the applied electrical as well as magnetic loads. This is not in agreement of the experimental findings. Therefore, GERR cannot be considered as a fracture parameter. It is also pointed that GERR is lower for (EP and MP) case.

The LERR is plotted with respect to λ_D and λ_B for different electromagnetic crack-face boundary conditions in Figs. 3(a) and 3(b), respectively. Figures 3(a) and 3(b) show that the LERR always increases by increasing electrical loads, λ_D, and magnetic loads, λ_B, respectively. It is in agreement with the experimental findings that negative and positive electrical and magnetic loadings always produce a shielding and unshielding effect on crack growth, respectively. It is also noted from Fig. 3(a) that LERR is uniform for the case (EP and MI) and (EP and MP). And from Fig. 3(b) that LERR is uniform for the case (EI and MP) and (EP and MP). Same as GERR, LERR is lower for (EP and MP) case too.

4 Conclusions

From the numerical studies, it is concluded that GERR cannot be considered as a fracture parameter because GERR is not confirming the experimental evidence. And also the numerical case study affirm that considered different crack models are capable to crack arrest for electrically and MP case under small-scale electromagnetic yielding.

References

1. Wang, B., Mai, Y.: Applicability of the crack-face electromagnetic boundary conditions for fracture of magnetoelectroelastic materials. Int. J. Solids Struct. **44,** 387–398 (2007)
2. Gao, H., Zhang, T.Y., Tong, P.: Local and global energy release rates for an electrically yield crack in piezoelectric ceramics. J. Mech. Phy. Solids **45,** 491–510 (1997)
3. Shen, S., Nishioka, T., Kuang, Z.B., Liu, Z.: Nonlinear electromechanical interfacial fracture for piezoelectric materials. Mech. Mater. **32,** 57–64 (2000)
4. Ma, P., Feng, W.J., Su, R.K.L.: An electrically impermeable and magnetically permeable interface crack with a contact zone in a magnetoelectroelastic bimaterial under uniform magnetoelectromechanical loads. Eur. J. Mech. A Solids **32,** 41–51 (2012)

Adaptive Matrix Transpose Algorithms for Distributed Multicore Processors

John C. Bowman and Malcolm Roberts

Abstract An adaptive parallel matrix transpose algorithm optimized for distributed multicore architectures running in a hybrid OpenMP/MPI configuration is presented. Significant boosts in speed are observed relative to the distributed transpose used in the state-of-the-art adaptive FFTW library. In some cases, a hybrid configuration allows one to reduce communication costs by reducing the number of message passing interface (MPI) nodes, and thereby increasing message sizes. This also allows for a more slab-like than pencil-like domain decomposition for multidimensional fast Fourier transforms (FFT), reducing the cost of, or even eliminating the need for, a second distributed transpose. Nonblocking all-to-all transfers enable user computation and communication to be overlapped.

1 The Matrix Transpose

The matrix transpose is an essential primitive of high-performance parallel computing. In contrast to the situation on serial and shared-memory parallel architectures, where the use of memory strides in linear algebra and fast Fourier transform (FFT) libraries allows matrices to be accessed in transposed order, many distributed computing algorithms rely on a global matrix transpose. This requires so-called all-to-all communication, where every process must communicate with all of the other processes to swap each matrix column with its corresponding row. For example, multidimensional FFT algorithms use a matrix transpose to localize the computation within individual processes. For efficiency, all data corresponding to a given direction must be made available locally for processing with the divide-and-conquer subdivision strategy of the FFT.

J. C. Bowman (✉)
University of Alberta, Edmonton, AB T6G 2G1, Canada
e-mail: bowman@ualberta.ca

M. Roberts
University of Strasbourg, Strasbourg, France
e-mail: malcolm.i.w.roberts@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_14

Writing an efficient implementation of a matrix transpose is surprisingly difficult. Even on serial and shared-memory machines there are implementation issues. While the storage savings afforded by in-place matrix transposition is often desirable, in-place matrix transposition on a serial machine is nontrivial for nonsquare matrices. For example, transposing $\begin{bmatrix} 0 & 1 & 2 & 3 \\ 4 & 5 & 6 & 7 \end{bmatrix}$ requires that the elements, stored linearly in memory, be permuted according to the cycles (0), $(1,4,2)$, $(3,5,6)$, and (7).

Algorithms for out-of-place matrix transposition are much simpler. Nevertheless, efficient implementation of out-of-place transposes still requires detailed knowledge of the cache size and layout, unless a recursive cache-oblivious algorithm is used [1]. For a review of serial in- and out-of-place matrix transposition algorithms, see [2].

On distributed memory architectures, a number of different matrix transposition algorithms have been proposed. For instance, Choi et al. [3] identified, in order of increasing speed, the *rotation*, *direct communication*, and *binary exchange* algorithms. However, the relative performance of these transposition algorithms depends on many factors, including communication latency, bandwidth, network congestion, packet size, local cache size, and network topology. Since it is hard to estimate the relative importance of these factors at compilation time, an adaptive algorithm, dynamically tuned to take advantage of these specific architectural details, is desirable. Al Na'Mneh et al. [4] have previously described an adaptive transposition algorithm for symmetric multiprocessors that share a common memory pool and exhibit low-latency interprocess communication. At the other extreme are adaptive algorithms optimized for distributed memory architectures with high-latency communication, like those implemented in the widely used FFTW library [5].

Modern high-performance computer architectures consist of a hybrid of the shared and distributed paradigms: distributed networks of multicore processors. The hybrid paradigm marries the high bandwidth low-latency interprocess communication featured by shared-memory systems with the massive scalability afforded by distributed computing. In this chapter, we describe recent efforts to exploit modern hybrid architectures, using the popular message passing interface (MPI) to communicate between distributed nodes and the OpenMP multithreading paradigm to communicate between the individual cores of each processor.

One of the obvious advantages of exploiting hybrid parallelism is the reduction in communication relative to the pure-MPI approach since messages no longer have to be passed between threads sharing a common memory pool. Another advantage is that some algorithms can be formulated, through a combination of memory striding and vectorization, so that local transposition is not required within a single MPI node (e.g., the multidimensional FFT[1]). The hybrid approach also allows smaller problems to be distributed over a large number of cores. This is particularly advantageous for 3D FFTs: the reduced number of MPI processes allows for a more slab-like than pencil-like domain decomposition, reducing the cost of, or even eliminating the

[1] However, the recent availability of serial cache-oblivious in-place transposition algorithms in some cases tips the balance in favor of local transposition, if transposed output is acceptable.

need for, a second transpose. A final reason in favor of the hybrid paradigm is that it is compatible with the modern trend of decreasing memory/core: the number of cores on recent microchips is growing faster than the total available memory. This restricts the memory available to individual pure-MPI processes.

Since the multicore nodes in modern hardware are typically connected to the distributed network via a single socket, message passing typically does not directly benefit from multithreading. However, we show in this chapter that message passing can benefit from the increased communication block lengths associated with the hybrid model. In addition, the necessary local transposes in and out of the communication buffer can benefit somewhat from multithreading.

The most popular algorithms for transposing an $N \times N$ matrix distributed over P processes are the *direct communication* (all-to-all) and recursive *binary exchange* (butterfly) algorithms. Direct communication transmits each block of data directly to its final destination in the matrix transpose, without any intermediate steps. It is most efficient for $P \ll N$, when the message sizes are large. However, its performance degrades for $P \approx N$, when the message size N^2/P^2 becomes small. To avoid this degradation, the binary exchange algorithm first groups messages together to reduce communication latency, by recursively subdividing the transpose into smaller block transposes.

The FFTW [5] library contains algorithms for both direct communication and binary exchange.

However, the FFTW implementation of an adaptive matrix transpose has been optimized for distributed memory architectures with high-latency communication. It does not effectively exploit the larger communication block sizes that are available with hybrid decompositions. It is also not multithreaded.

We have developed an efficient hybrid algorithm in the open-source library FFTW++ [6]. It uses direct communication when the message sizes are large and a two-stage block transpose in latency bound cases. In the latter case, we divide the total number of processes P into a blocks each containing b processes. A block transpose expresses an $N \times M$ matrix as an $a \times a$ matrix of $N/a \times M/a$ blocks. Here, we only discuss the case where $P = ab$ divides N and M; the general case can be handled with minor technical modifications. The transpose of each $N/a \times M/a$ blocks is computed first, followed by the transpose of the $a \times a$ matrix of blocks. Grouping is used to increase the message sizes in the first transpose from NM/P^2 to aNM/P^2.

The binary exchange algorithm performs recursive block transposes. In practice, only one level (at most) of recursion is actually necessary to increase the communication message sizes. After that single recursion, direct communication typically becomes optimal since the message sizes have now been multiplied by a factor of a in the first phase and b in the second phase. We show theoretically in Sect. 2 that the communication costs are minimized for $a = b = \sqrt{P}$. In practice, the optimal value will lie somewhere near this value, but may vary due to other considerations, such as cache configuration and network topology.

Block transposition is illustrated for the case $N = M = 8$, $a = 4$, and $b = 2$ in Fig. 1. In (a), the transpose of each 2×2 block is computed. The communications

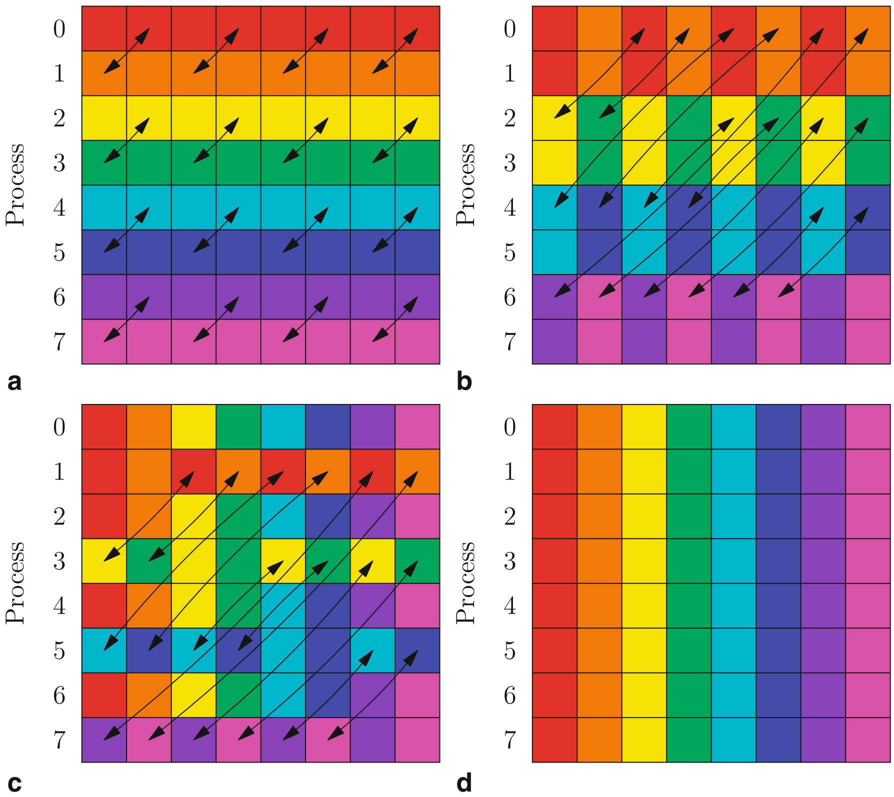


Fig. 1 An 8×8 block transpose over eight processes for the case $a = 4$ and $b = 2$

between pairs $(2n, 2n+1)$ of processes are grouped together by first doing an out-of-place local transpose of the data, considered as a 4×2 matrix, on each process. The pairs of processors then exchange data, as indicated by the arrows. This is followed by separate all-to-all communications between the even processes (b) and odd processes (c), again grouping the data bound for identical processors, to obtain the transposed matrix in (d).

The block transposition algorithm may be stated as follows:

1. Inner transpose:
   a. Locally transpose $N/b \times b$ matrix of blocks of M/P elements.
   b. All-to-all communicate over teams of b processes, using block size aNM/P^2.
2. Outer transpose:
   a. Locally transpose $N/a \times a$ matrix of blocks of M/P elements.
   b. All-to-all communicate over teams of a processes, using block size bNM/P^2.
3. Locally transpose $N \times M/P$ matrix (optional).

Step 2 is omitted when $a = 1$ (the nonlatency bound case); the algorithm then reduces to direct communication. Step 3 can be omitted if local transposition of the output data is not required. We designed our algorithm to use nonblocking communications (MPI_Ialltoall, available in MPI 3.0), to allow the user to overlap computation with one or even both communication phases. Overlapping computation with communication has been observed to yield modest speedups (roughly 10 %) for computing 3D implicitly dealiased convolutions [6, 7], where a natural parallelism between communication and computation arises.

2 Communication Costs

Direct transposition of an $N \times M$ matrix distributed over P processes, involves $P-1$ communications per process, each of size NM/P^2, for a total per-process data transfer of $(P-1)NM/P^2$. For large P, this cost asymptotically approaches NM/P.

For a block transpose, one exploits a factorization $P = ab$ to perform the transform in two stages. First, one groups the processes into a teams of b according to the quotient of their rank and b. Over each team of b processes, one computes the inner transpose of the a individual $N/a \times M/a$ matrices, grouping all a communications with the same source and destination together. This requires $b-1$ messages per process, each of size $(NM/a)/b^2 = aNM/P^2$, for a total per-process data transfer of $(b-1)aNM/P^2$. One then regroups the processes into b teams of a according to their rank modulo b. Over each team of a processes, the outer transpose of the $a \times a$ matrix of $N/a \times M/a$ blocks requires $a-1$ communications per process, each of size $(NM/b)/a^2 = bNM/P^2$, for a total per-process data transfer of $(a-1)bNM/P^2$.

Each process performing a block transpose must therefore send $(a-1)+(b-1) = a + P/a - 2$ messages, for a total per-process transfer of

$$[(b-1)a + (a-1)b]\frac{NM}{P^2} = \left(2P - a - \frac{P}{a}\right)\frac{NM}{P^2}.$$

Let τ_ℓ be the typical latency of a message and τ_d be the time required to send each matrix element. The time required to perform a direct transpose is

$$T_D = \tau_\ell\,(P-1) + \tau_d\frac{P-1}{P^2}NM = (P-1)\left(\tau_\ell + \tau_d\frac{NM}{P^2}\right),$$

whereas a block transpose requires

$$T_B(a) = \tau_\ell\left(a + \frac{P}{a} - 2\right) + \tau_d\left(2P - a - \frac{P}{a}\right)\frac{NM}{P^2}.$$

Since

$$T_D - T_B = \tau_d\left(P + 1 - a - \frac{P}{a}\right)\left(L - \frac{NM}{P^2}\right),$$

where $L = \tau_\ell/\tau_d$ is the effective communication block length, we see that a direct transpose is preferred when $NM \geq P^2L$, while a block transpose should be used

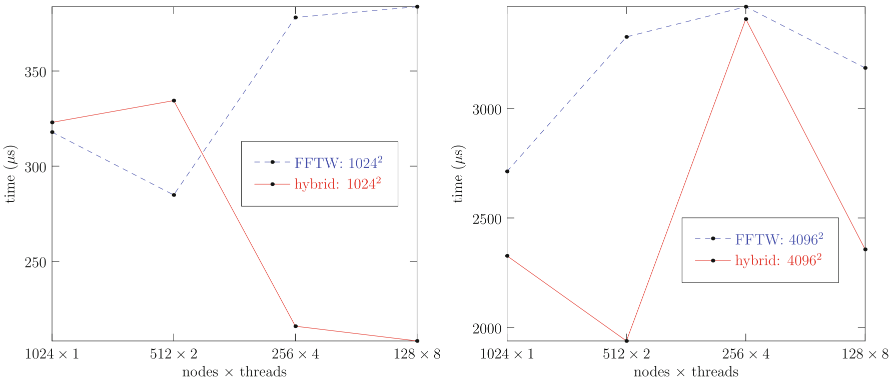


Fig. 2 Wall-clock times for distributed transposes with the fastest Fourier transform in the west (FFTW) library vs. our implementation

when $NM < P^2L$. To determine the optimal value of a for a block transpose, consider

$$T_B'(a) = \tau_\ell\left(1-\frac{P}{a^2}\right)+\tau_d\left(-1+\frac{P}{a^2}\right)\frac{NM}{P^2} = \tau_d\left(1-\frac{P}{a^2}\right)\left(L-\frac{NM}{P^2}\right).$$

For $NM < P^2L$, we see that T_B is convex, with a global minimum value at $a=\sqrt{P}$ of

$$2\tau_d\left(\sqrt{P}-1\right)\left(L+\frac{NM}{P^{3/2}}\right) \sim 2\tau_d\sqrt{P}\left(L+\frac{NM}{P^{3/2}}\right), \qquad P \gg 1.$$

The global minimum of T_B over both a and P is then seen to occur at $P \approx (2NM/L)^{2/3}$. If the matrix dimensions satisfy $NM > L$, as is typically the case, this minimum occurs above the transition value $(NM/L)^{1/2}$. For $P \gg 1$, we note that $T_D \sim \tau_d(PL + NM/P)$ has a global minimum of $2\tau_d(NML)^{1/2}$ at $P = (NM/L)^{1/2}$, precisely at the transition between the two algorithms. Provided $NM > L$, the optimal choice of P is thus $(2NM/L)^{2/3}$. On a multicore parallel run over S sockets, with C cores per socket, the optimal number of OpenMPI threads to use is then $T = \min(SC/(2NM/L)^{2/3}, C)$, with $P = SC/T$ MPI nodes. We benchmarked our hybrid implementation against the FFTW transpose for 1024×1024 and 4096×4096 complex matrices on the Dell Zeus C8220 Cluster at the Texas Advanced Computer Center, using $S = 128$ sockets and $C = 8$ cores. In Fig. 2, we see that our implementation typically outperforms FFTW, in some cases by nearly a factor of 2. We measured the value of L to be roughly 4096 bytes for this machine. This predicts that the optimal number of threads is $T = 8$ for Fig. 2a and $T = 2$ for Fig. 2b, precisely as observed.

Acknowledgments The authors gratefully acknowledge Professor Wendell Horton for providing access to state-of-the-art computing facilities at the Texas Advanced Computer Center.

References

1. Frigo, M., Leiserson, C.E., Prokop, H., Ramachandran, S.: Foundations of Computer Science, 1999. 40th Annual Symposium on (IEEE, 1999), pp. 285–297
2. Dow, M.: Transposing a matrix on a vector computer. Parallel Comput. **21**(12), 1997 (1995)
3. Choi, J., Dongarra, J.J., Walker, D.W.: Parallel matrix transpose algorithms on distributed memory concurrent computers. Parallel Comput. **21**(9), 1387 (1995)
4. Al Na'mneh, R., Pan, W.D., Yoo, S.M.: Efficient adaptive algorithms for transposing small and large matrices on symmetric multiprocessors. Informatica **17**(4), 535 (2006)
5. Frigo, M., Johnson, S.G.: The design and implementation of FFTW3. Proc. IEEE **93**(2), 216 (2005)
6. Bowman, J.C., Roberts, M.: FFTW++: A fast Fourier transform C^{++} header class for the FFTW3 library. http://fftwpp.sourceforge.net (2010)
7. Bowman, J.C., Roberts, M.: SIAM J. Efficient dealiased convolutions without padding, SIAM. Sci. Comput. **33**(1), 386 (2011)

Accounting for Temperature when Modeling Population Health Risk Due to Air Pollution

Wesley S. Burr and Hwashin H. Shin

Abstract Air Health Indicator (AHI) is a joint Health Canada/Environment Canada initiative. A component in the indicator is an estimate of the time-dependent population health risk due to short-term (acute) effects of air pollution. The standard approach for this risk estimation uses a generalized additive model (GAM) framework, which includes one or more air pollutants and one or more temperature terms as covariates, as well as a smooth function of time. In this risk-modeling framework, the temperature is not the primary focus, but is included to ensure that common structure between the mortality (response), the pollutant(s), and the temperature is not included in the risk attribution.

We examine the smooth function link that is commonly used when including temperature. We show that for a single lag of temperature, the traditional J-, U-, or V-shaped relationship between temperature and mortality is largely a function of low-frequency mortality structure and is thus accounted for by the smooth function of time typically included in risk models. We further compare and contrast the first two primary lags of temperature in the context of these findings, and demonstrate differences in their structure, advocating the inclusion of only the first (lag-0) parametric temperature series in the model.

1 Introduction

Health Canada has recently developed a new methodology, the Air Health Indicator (AHI), for assessing the effects on daily mortality of short-term exposure to air pollution as they may vary dynamically over space and time in response to changes in air quality. Hundreds of time-series studies of daily mortality have now been published worldwide, and are critical components of the scientific evidence supporting

W. S. Burr (✉) · H. H. Shin
Queen's University, Kingston, ON, Canada
e-mail: wburr@mast.queensu.ca

H. H. Shin
e-mail: hhshin@rogers.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_15

a causal relationship between air pollution and public health. The AHI provides time trends in annual risks at city-specific, regional, and national levels such as an increasing, decreasing, or constant trend over a time period. The AHI can be used in policy analysis, with potentially important applications to the assessment of the public health impacts of air quality regulation. The AHI is computed by using a standard generalized additive model (GAM) framework, of which temperature is an integral part.

The GAM formulation for estimation of risk due to air pollution includes temperature as a predictor as it is known to be one of the strongest (and quite clearly causal) predictors of daily mortality. Since the classic response relationship [4, 9] between mortality and temperature is concave up (i.e., *J*-, *U*-, or *V*-shaped), this has been the impetus behind including temperature as a nonparametric-smoothed predictor. However, after accounting for long-timescale variation in the model, the response curve changes behavior dramatically to be approximately monotonically increasing. Additionally, when considering temperature, a common concern (see, e.g., [1]) is determining the most appropriate lag (or lags) for the included term(s). Considering the previous point, we examine the interplay between the response and parametric temperature and add slightly to the evidence for lag-0 temperature alone being the sensible choice.

2 Model Used

For the purposes of the AHI, the model used links the response (mortality counts) to the air pollutant predictor of interest (e.g., Ozone, NO_2) via an additive model structure. Specifically, a GAM is used with a Poisson or quasi-Poisson family assumed. This family has a logarithmic function link, with all predictors entering additively. The predictors used are a single air pollutant, one or more temperature terms, a day-of-week (DOW) factor term, and a smooth function of time. The smooth function of time is included to remove slowly varying long-timescale structure from the response, and to control the autoregressive relationship that is inherent in a time series of observations. Formally,

$$\log(\mu_t) = \gamma_0 \mathbf{x}_t + \gamma_1 \text{DOW} + \sum_{j=1}^{K} \beta_j S_j(T_{j,t}, \text{dof} = d_j) + S_{K+1}(\text{time}, \mathit{dof} = 14/\text{year}) \tag{1}$$

where the $S_j(\cdot)$, $j = 1, \cdots, K$ are possible identity links, and S_{K+1} is a cubic regression spline smoother. Note that the typical choice of *dof* for such a smooth function of time is 7/year, an unfortunate misunderstanding that has crept into the literature—to correctly remove structure of longer timescale than 7 cycles/year, it is necessary to use 14 *dof*/year instead [11]. The DOW is the day-of-week factor term, $\mathbf{x}_t$ is the pollutant of interest, and $T_{j,t}$ is the jth temperature term, typically consisting of some combination of separate lags of daily mean temperature.

2.1 Previous Understanding

Previously published works (see especially [4, 10]) have emphasized the necessity of including temperature via a smooth function link. In the words of Dominici et al. [4] (p. 268):

> … we also fitted smooth functions of the same day temperature (temp$_0$), the average temperature for the three previous days (temp$_{1-3}$), each with 6 degrees of freedom … [in] US cities, mortality decreases smoothly with increases in temperature until reaching a relative minimum and then increases quite sharply at higher temperature. 6 degrees of freedom were chosen to capture the highly non-linear bend near the relative minimum as well as possible.

This behavior can clearly be seen in Fig. 1, where temperature at various lags is shown fit to all-cause daily mortality via cubic regression splines with the specified *dof*. All models are fit using the `R` programming language [8] using the `mgcv` [12] and `spsmooth` [3] packages. As clearly seen in nine panels of this figure, the nonlinear "bend" mentioned by [4] is present. However, this figure does not tell the whole story, as this relationship is between *raw* log mortality and daily mean temperature (at some lag). When considered in the context of a larger population health risk model, we must remember that temperature is not fit to the raw log mortality, but rather to the filtered log mortality (via the influence of the smooth function of time). We consider this in the next section.

2.2 Results Inside the Model

Rather examining the relationship between raw log mortality and temperature in order to determine how best to include the temperature in the model, we expand the context slightly and consider the presence of the smooth function of time. When considered from a signal processing or time series point of view, the smooth function of time is (when included as a fixed-*dof* spline) quite simple: a linear filter. Linear filters are particularly easy to understand, and in this case can be interpreted as capturing a portion of the variation of the response, leaving the residual to be fit by the temperature term via its function link. If 14 *dof*/year are used, as discussed previously in this section, then effectively the filter (or smooth function of time) will capture the bulk of the power in the response that varies on timescales longer than 7 cycles/year, or roughly 52 days. Then, the temperature will be smoothed against the residual from this filtering operation, or equivalently, the power in the response that varies on timescales shorter than 7 cycles/year. The results of this can be seen in Fig. 2, where a clear change is observed.

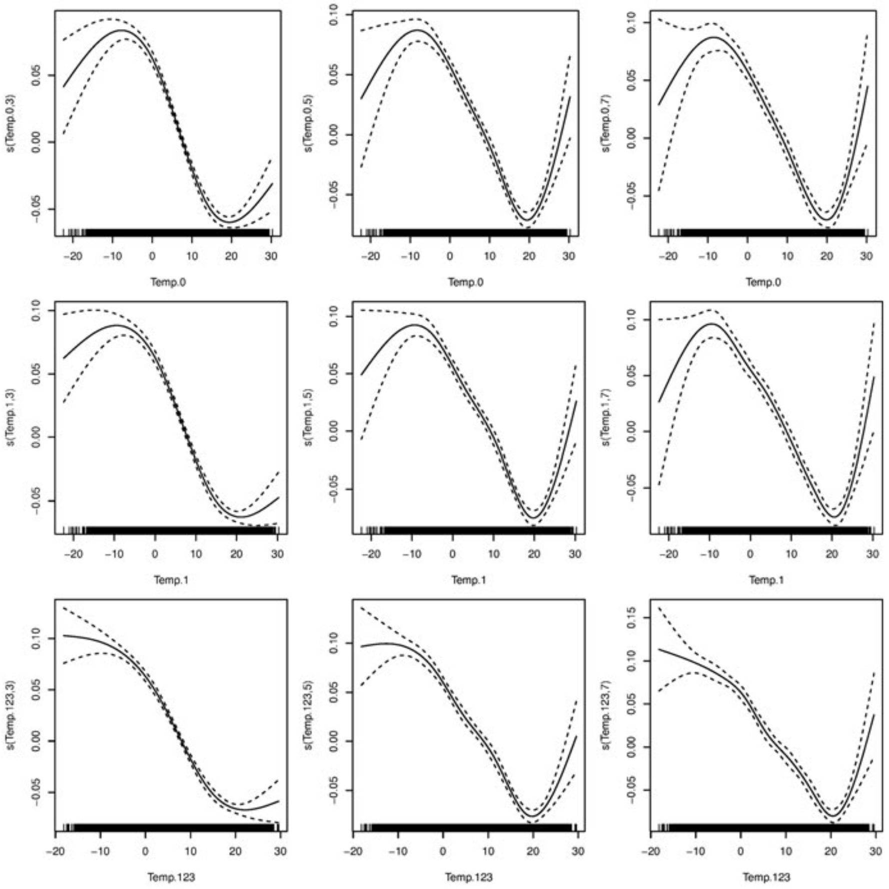


Fig. 1 Nine smooths of daily mean temperature against daily all-cause mortality are shown for the city of Toronto, Ontario using data from 1981 through 2007. Going down in rows, each row represents a particular choice of temperature term: lag-0, lag-1, and the averages of lags 1, 2, and 3. Going across, each column represents a different choice of *dof* for the cubic regression spline smoother function link. The model used was a simple single-predictor GAM with Poisson link to simulate the logarithmic function link that is used in the full health risk model

2.3 Interpretation

The plots from the previous two sections show a clear change when the smooth function of time is taken under consideration. Accordingly, the rationale for including temperature via a smooth function link is somewhat suspect. After all, if the justification is that we wish to capture a nonlinear bend in the relationship between the two, and the nonlinear bend effectively disappears when we consider the model properly in context, then the justification is moot. The approximately monotonic relationship shown in the top six panels of Fig. 2 imply that including temperature in

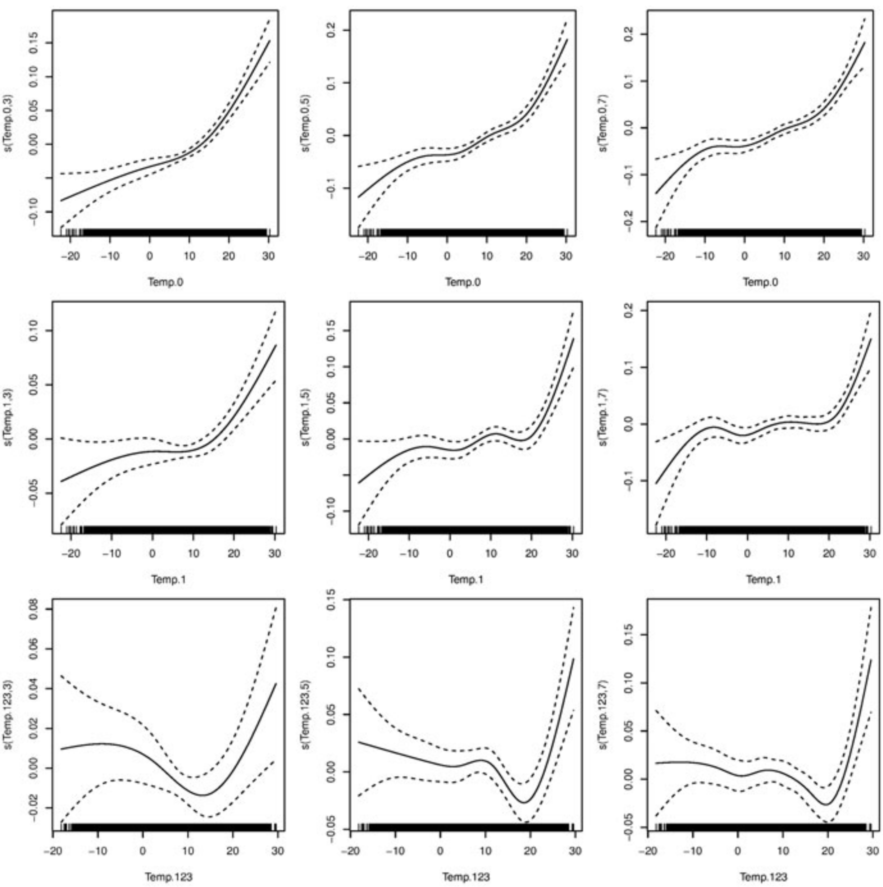


Fig. 2 Nine smooths of daily mean temperature against daily all-cause mortality are shown for the city of Toronto, Ontario using data from 1981 through 2007. All layout concerns are identical to Fig. 1, but the model linking the two terms has an additional smooth function of time added

the model parametrically may be equivalent (as a parametric inclusion corresponds to a smooth function link with 1 *dof*, i.e., a straight line). Further examination of the larger scale model of Eq. (1) as applied to 24 large urban centers in Canada (with the AHI dataset) show that the risk estimates obtained using smooth function links with 3 *dof* are comparable and share similar structure to estimates obtained using a simple parametric inclusion. As parametric terms are simpler (parsimony) and allow for more straightforward interpretation, these results lead us to advocate for their inclusion to be parametric for the AHI, and similar metrics.

3 Considering Different Lags

The previous section showed that including temperature via a smooth function link is unnecessary when also including a smooth function of time (filter) term in additive models such as we are considering for the AHI. What remains uncertain is what this finding implies for the choice of *lag* for temperature. Previous models have used multiple terms so as to capture as much of the temperature–mortality relationship as possible. Examples include the additive combination of lag-0 and lag-1 (AHI), lag-0 and the average of lags 1, 2, and 3 (Dominici et al. [4, 7]), and the more sophisticated distributed-lag nonlinear model approach of Gasparrini and Armstrong [5, 6].

For sake of brevity, we will consider only the AHI form of including lags 0 and 1 as separate terms in the model. While this form was chosen via standard metrics (e.g., Akaike Information Criterion (AIC)), previously there has been little work done on considering the effect of including the second term in the model. Note that by "second term," we acknowledge that any reasonable model of air pollution and health effects must include lag-0 temperature as a critical component in the model, as it is the most efficacious of any possible choice. Note that the findings of Gasparrini [5] suggest that while temperature and mortality can be related out to lag 20 days and higher, the majority of the relationship occurs in the first three to four lags (i.e., lags 0 to 3).

Now, consider Fig. 3, in which we show the prediction (coefficient multiplied by series) of lag-0 and lag-1 temperature for Toronto, again between 1981 and 2007. The two predictions are almost perfectly in-phase. This is a curious result, as a roughly equivalent effect can be obtained by simply including lag-0 alone with a slightly larger coefficient.

We compared the mean risk (coefficient of the air pollutant of interest) across 24 Canadian urban centers for models with only lag-0 temperature, only lag-1, no temperature at all, and both lag-0 and lag-1. In order, from the lowest average risk magnitude to highest, the choices are: lag-0 alone, lag-0, and lag-1 (comparable to lag-0), then lag-1, and finally no temperature term at all. The lag-1 models have risks that are comparable to the no-temperature term models, while the lag-0 and lag-1 models together are marginally higher than the lag-0. These results indicate that by including temperature lag-1 together in a model with lag-0, the temperature term actually captures *less* of the variation in the mortality, resulting in slightly higher average coefficients for the air pollution term. Thus, from the point of view that says that we should account for as much of the variation in the mortality as possible before attributing the residual to air pollution, the best model to use, drops temperature lag-1 entirely, leaving only a parametric inclusion of temperature lag-0.

4 Conclusion

A reconsideration of the inclusion of temperature in models of population health risk due to air pollution shows that when considered in the context of the overall model, the smooth function link used for temperature lacks rationale for its use. Accordingly,

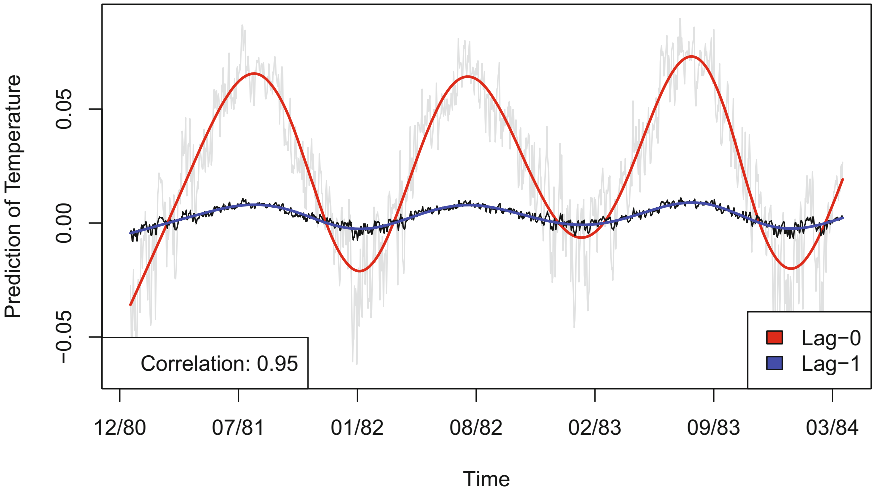


Fig. 3 Comparison of the predictions for lag-0 and lag-1 temperature for Toronto between 1981–2007. The model fit is in the form of Eq. 1 with $K = 2$, and the temperature terms are included parametrically. The correlation between the two temperature predictions is 0.95, indicating that they are almost perfectly in-phase with one another

models may include temperature parametrically, leading to the increased intuition and ease of implementation. Considering the first two lags of temperature, we have further demonstrated that, for Canadian urban centers, lag-1 seemingly adds no value to models which also include lag-0. This result suggests that mortality for Canada is driven by near-instantaneous temperature, with recent historical exposure being insignificant in comparison.

Acknowledgement The authors gratefully acknowledge discussion with Professor Glen Takahara, Queen's University, and data made available by the Air Health Indicator (AHI) project. Some content has been previously published as a part of the PhD thesis of the first author [2].

References

1. Barnett, A.G., Tong, S., Clements, A.C.A.: What measure of temperature is the best predictor of mortality? Environ. Res. **110**(6), 604–611 (2010)
2. Burr, W.S.: Air Pollution and Health: Time Series Tools and Analysis. PhD thesis, Queen's University, Kingston, Ontario, Canada, October 2012
3. Burr, W.S., contributions from Karim Rahim: spsmooth: An Extension Package for 'mgcv', (2012). R package version 1.0–0
4. Dominici, F., Samet, J.M., Zeger, S.L.: Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. J. Royal Stat. Soc. Series A Stat. Soc. **163**(3), 263–302 (2000)
5. Gasparrini, A., Armstrong, B., Kenward, M.G.: Distributed lag non-linear models. Stat. Med. **29**(21), 2224–2234 (2010)

6. Goldberg, M.S., Gasparrini, A., Armstrong, B., Valois, M.-F.: The short-term influence of temperature on daily mortality in the temperate climate of Montreal, Canada. Environ. Res. **111**(6), 853–860 (2011)
7. Peng, R.D., Dominici, F., Louis, T.A.: Model choice in time series studies of air pollution and mortality. J. Royal Stat. Soc. Series A Stat. Soc. **169**(2), 179–203 (2006)
8. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2012). ISBN 3-900051-07–0
9. Samet, J., Zeger, S., Kelsall, J., Xu, J., Kalkstein, L. Does weather confound or modify the association of particulate air pollution with mortality? An analysis of the philadelphia data, 1973–1980. Environ. Res. **77**(1), 9–19 (1998)
10. Schwartz, J., Spix, C., Touloumi, G., Bachárova, L., et al.: Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. J. Epidemiol. Community Health **50**(Suppl 1), S3–11 (1996)
11. Slepian, D.: Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: the discrete case. Bell Syst. Tech. J. **57,** 1371–1429 (1978)
12. Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. Royal Stat. Soc. Series B Stat. Methodol. **73**(1), 3–36 (2011)

Discrete Prolate Spheroidal Sequences as Filters in Generalized Additive Models

Wesley S. Burr and Hwashin H. Shin

Abstract Air Health Indicator (AHI) is a joint Health Canada/Environment Canada initiative that seeks to model the Canadian national population health risk due to short-term (acute) effects of air pollution. The commonly accepted model in the field uses cubic spline-based temporal smoothers embedded in generalized additive models (GAMs) to account for seasonal and long-term variations in the response. From a spectral point of view, it is natural to think of these smooth, long-term variations as low-frequency components, and the temporal smoother as a linear filter.

Examining the frequency response of the filters typically used, we show that the performance leaves much to be desired. Adapting the discrete prolate spheroidal sequences as filters, taking inspiration from their similar use in the multitaper method, we are able to significantly improve the frequency response of the smoother. We conclude with a discussion of the implications for controlling bias from the long timescale structure of parametric covariates, and suggest a prefiltering stage to such models.

1 Introduction

Health Canada has recently developed a new methodology, the Air Health Indicator (AHI), for assessing the effects on daily mortality of short-term exposure to air pollution as they may vary dynamically over space and time in response to changes in air quality. Hundreds of time-series studies of daily mortality have now been published worldwide, and are critical components of the scientific evidence supporting a causal relationship between air pollution and public health. The AHI provides time trends in annual risks at city-specific, regional and national levels such as an increasing, decreasing, or constant trend over a time period. The AHI can be used in policy analysis, with potentially important applications to the assessment of the public health

W. S. Burr (✉) · H. H. Shin
Queen's University, Kingston, ON, Canada
e-mail: wburr@mast.queensu.ca

H. H. Shin
e-mail: hhshin@rogers.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_16

impacts of air quality regulation. The AHI is computed by using a standard generalized additive model (GAM) framework, of which a smooth function of time (linear filter) is an integral part.

2 Model Used

For the purposes of the AHI and similar estimates of population health risk (e.g., NMMAPS [4, 9]) the model used links the response (mortality counts) to the air pollutant predictor of interest (e.g., Ozone and NO_2) via an additive model structure. Specifically, a GAM is used with a Poisson or quasi-Poisson family assumed. This effectively sets the functional link to be logarithmic, with all predictors entering additively. The predictors used are a single air pollutant, one or more temperature terms, a day-of-week factor term, and a smooth function of time. The smooth function of time is included to remove slowly varying long timescale structure from the response, and to control the autoregressive relationship that is inherent in a time series of observations. Formally:

$$\log(\mu_t) = \gamma_0 \mathbf{x}_t + \gamma_1 \mathrm{DOW} + \sum_{j=1}^{K} \beta_j S_j(T_{j,t}, \mathrm{df} = d_j) + S_{K+1}(\mathrm{time}, df = 14/\mathrm{year}), \tag{1}$$

where the $S_j(\cdot)$, $j = 1, \ldots, K$ are possible identity links, and S_{K+1} is a cubic regression spline smoother. Note that the typical choice of degrees-of-freedom (*dof)* for such a smooth function of time is 7/year, an unfortunate misunderstanding that has crept into the literature. We will discuss this further below. The DOW is the day-of-week factor term, $\mathbf{x}_t$ is the pollutant of interest, and $T_{j,t}$ is the jth temperature term, typically consisting of some combination of separate lags of daily mean temperature. The notation used is similar to that of Dominici et al. [2], and the models are implemented and computed in the `R` [8] programming language.

2.1 Rationale for Including the Smooth Function of Time

The smooth function of time is traditionally included in these models [2, 5] to account for unmeasured confounding. Mortality has a strong seasonal variation, which in Canada and the USA is largely driven by the seasons of the continental and subtropical climates. There are any number of causal factors that contribute to nonaccidental cardiovascular or cardiopulmonary mortality, many of which are not measured routinely, or in some cases, easily. In many of these causal risk factors, long timescale structure is also present, e.g., influenza epidemic cycles peak in midwinter for the northern hemisphere, and demographic shifts vary on multiyear or decadal time scales. Accordingly, the standard model includes a smooth function of time to account for these unmeasured factors. Additionally, the smooth function of time is included so as to account for any additional temporal correlation in the log-mortality count series. The intention is to allow for the risk (coefficient of the pollutant) to be estimated using only short-term variations in mortality and air pollution.

In practice, this inclusion effectively filters the response, capturing much of the variation in the low frequency range, leaving only high frequency structure. Of course, the bandwidth of the filter is depends upon the choice of smoother, which in the commonly accepted model is 6 or 7 *dof*/year. Unfortunately, the prior information used to select this *dof* value suggests that the bandwidth should account for the variation at periods 6 weeks and greater, yet the choice of 7 *dof*/year for a cubic regression spline smoother equates [11] to almost twice that level: periods of approximately 4 months and greater.

3 Transfer Functions: Splines and Prolates

In the default model, cubic regression splines with a fixed number of basis vectors (equating to *dof*) are primarily used due to issues [7] with concurvity, the nonparametric analogue of multicollinearity. As an example, consider a model using 10 years of data with a smooth function of time chosen to have 7 *dof*/year (70 *dof* total). This equates to a smoother matrix with 70 basis vectors, and is thought [3, 2, 5] (among many others) to capture variation corresponding to periods longer than 7 cycles/year (52.2 days). Unfortunately, 70 basis functions actually corresponds [10, 11] to roughly $W = 70/2N$ or 0.00958 Hz, a period of roughly 100 days. Thus, the actual bandpass of this smoother is not what was intended, and further, the magnitude response performance of the filter is also decidedly suboptimal (see Fig. 1).

Instead by using a filter composed of discrete prolate spheroidal sequences (*Slepians* for short, due to the contributions of David Slepian [10, 11]), significant improvements can be seen in both pass-band performance and out-of-band power suppression. This choice is based on the earlier work of Papoulis [6] in which modifications are presented for the default Slepian tapers. Furthermore, there is a direct link between a desired bandwidth W for the smoother and the number of basis vectors (hence, *dof*) required to adequately represent the subspace in question. Thus, by using a priori information and selecting a desired passband of periods 50 days and longer, (for our previously chosen example) we immediately deduce that no more than $M = 2\,\mathrm{NW} = 2 \cdot 10 \cdot 365 \cdot 1/50 = 146$ basis vectors will be required. Further, Slepian's work [11] indicates that $M \approx 2\,\mathrm{NW} - 2$ or 144 basis vectors can be chosen so that all are sufficiently concentrated in the appropriate passband. Note that this is approximately twice the number of *dof* thought necessary by previous work.

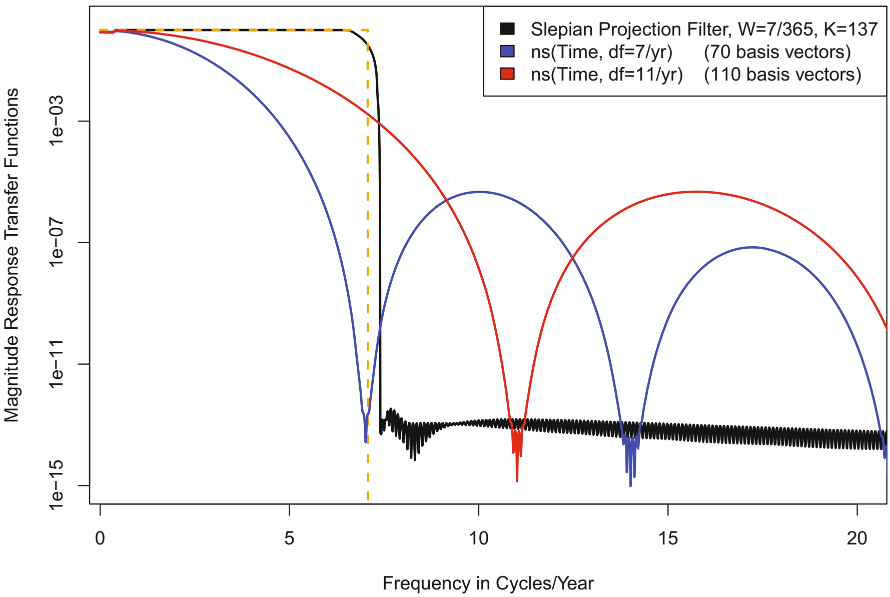


Fig. 1 Magnitude response transfer functions for two cubic regression spline smoothers (*dof* as indicated) as well as a similarly chosen discrete prolate spheroidal sequence smoother. Note the sharp band-edge for the prolate smoother, and the flat passband. One-hundred thirty-seven basis vectors were used for the prolate smoother, as $\lfloor 2\,NW \rfloor \approx 139$

4 Residual Effective Response: Internal Model Comparison

As mentioned above, included in the rationale for including a smooth function of time in models for estimation of population health risk is a justification for capturing long timescale variation in the response (typically log mortality). However, there is a misunderstanding in the literature which implies that more than this is inherently possible. Taken from [5]:

> ... the smooth function of time services as a linear filter on the mortality **and pollution series** and removes any seasonal or long-term trends in the data ...

It is difficult to see how this quote can imply anything but what it clearly says: that the smooth function of time acts as an effective filter not only on the response (by capturing the long timescale variation, it acts as a high-pass filter, leaving the short timescale variation untouched for the rest of the model) but *also* on the pollutant series, which is typically included parametrically. Unfortunately, while this statement may be true in a situation where the effective model-fitting paradigm caused the response to be filtered once and then never modified, all GAM fitting algorithms instead use a form of iteratively reweighted least squares, with emphasis on the *iteratively*. As such, due to the application of the filter to the response after accounting for the other predictors (which we call the *residual effective response*) the implied filtering effect on the pollutant is not fully realized.

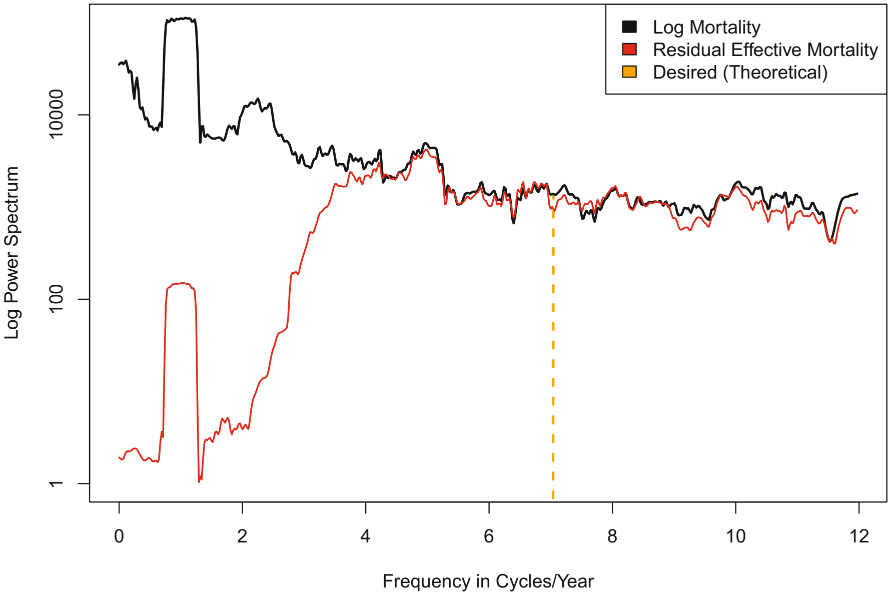


Fig. 2 Power spectrum estimates for both log mortality and the residual effective mortality as observed by the pollutant, in this case daily mean ozone, `o3tmean`. Data are taken from the `NMMAPS` database [4], and the results shown are computed using Chicago, 1987–2000. Note the large amount of structural power between 0 and 7 cycles/year, representing long timescale variation that the pollutant will be fit to as a portion of the regression. Part of this power is due to poor choice of time smoother (cubic regression splines with 7 *dof*/year), and part due to bleed-over from the long timescale portion of the primary covariate (ozone)

A typical pollutant in a population health risk model is a full-spectrum time series of daily measurements. Accordingly, each iteration of the GAM solver will apply the smooth function of time filter to a residual effective mortality consisting of a high-pass filtered series *plus* a scaled copy of the pollutant series, among possible others. The subsequent iterations will then filter this residual effective series, and no matter how well the filter may work, there will always be a portion of the residual effective response as observed by the *pollutant* that will contain a scaled copy of the pollutant itself—see Fig. 2 for a demonstration of this using an all-ages variant of the model of Dominici et al. ([2], p. 278). Simplified models which contain only a parametric pollutant term and a time filter suggest that the coefficient obtained by the model solver is scaled by $1/(1-r)$, where r is the percentage of variation in the pollutant below the chosen smooth function filter bandwidth W. Thus, to suggest that the smooth function of time acts as an effective linear filter on the pollutant is incorrect, although the underlying expectation is a useful goal. In the next section, we present an alternative which meets the stated goal.

5 Discussion and Conclusion

With some understanding of the behavior of the iterative solvers inherent in GAMs, we propose an alternative approach which meets the suggestion of the quote of Peng et al. [5] (above). Rather than trust the smooth function of time to account for all long timescale variation in the model (both in the response and pollutant), we propose applying a prior stage to the model, in effect capturing the long timescale variation in the pollutant *before* introducing it in the model. This can be done in a number of ways, including a second application of GAMs, but the easiest way is to simply apply a linear filter using the same construction as is used for creation of the model/basis matrix for a time smoother.

Applying this prior-stage filter to the pollutant series prevents any contamination from the long timescale structure of the same, and ensures that the coefficient obtained truly represents whatever a priori timescale was chosen. For example, using a bandwidth of 7 cycles/year results in both a residual effective mortality and a pollutant which simultaneously have little structure at periods longer than 52 days. This ensures that any iterative process used to estimate risk fits only the short timescale portion of the pollutant to a short timescale-focused residual effective response. The early results are quite promising, and while the process does add a computational burden (due to the extra stage), protecting risk estimates from bias due to long timescale pollutant bleed-over seems to be a worthy goal. We feel that completion of this work will correct for the issues noted in this chapter.

In summary, examining GAMs used for estimation of population health risk due to air pollution, we were able to show that a misunderstanding exists in the literature. To fully capture long timescale variation in the response, it is necessary to use a correct $\approx 2\,NW$ basis vectors for a time-based smoother. Further moving from a cubic regression spline smoother to a discrete prolate spheroidal sequence smoother results in improved pass-band performance and a sharper band-edge, with no loss in *dof* but with some increase in computational burden. Examining the estimation of models using the improved prolate time-based smoothers gives improved intuition regarding the interaction of parametric and filter terms in such models, and suggests that a prior-stage filtering for pollution covariates will reduce bias in coefficient estimates.

Acknowledgment The authors gratefully acknowledge discussion with Professor Glen Takahara, Queen's University. Some material contained in this chapter has been previously published as part of the PhD thesis of Wesley S. Burr [1].

References

1. Burr, W.S.: Air Pollution and Health: Time Series Tools and Analysis. PhD thesis, Queen's University, Kingston, Ontario, Canada, October (2012)
2. Dominici, F., Samet, J.M., Zeger, S.L.: Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. J. R. Stat. Soc. Ser. A (Statistics in Society) **163**(3):263–302 (2000)

3. Dominici, F., McDermott, A., Zeger, S.L., Samet, J.M.: On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. **156**(3):193–203 (2002)
4. Peng, R.D., Welty, L.J.: The NMMAPSdata package. R News **4**(2):10–14 (2004)
5. Peng, R.D., Dominici, F., Louis, T.A.: Model choice in time series studies of air pollution and mortality. J. R. Stat. Soc. Ser. A (Statistics in Society) **169**(2):179–203 (2006)
6. Papoulis, A., Bertran, M.S.: Digital filtering and prolate functions. IEEE Trans. Circuit Theory **19,** 674–681 (1972)
7. Ramsay, T.O., Burnett, R.T., Krewski, D.: The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology **14**(1):18–23 (2003)
8. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0. (2012)
9. Samet, J., Dominici, F., Zeger, S., Schwartz, J., Dockery, D.W.: National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and Methodologic Issues. Technical Report 94-I, Health Effects Institute, 2000. Part I of IV
10. Slepian, D.: Prolate spheroidal wave functions, Fourier analysis and Uncertainty: IV. Bell Syst. Tech. J. **43,** 3009–3057 (1964)
11. Slepian, D.: Prolate spheroidal wave functions, Fourier analysis, and Uncertainty V: the Discrete Case. Bell Syst. Tech. J. **57,** 1371–1429 (1978)

Time Series Analysis and Calibration to Option Data: A Study of Various Asset Pricing Models

Giuseppe Campolieti, Roman N. Makarov and Arash Soleimani Dahaj

Abstract In this chapter, we study three asset pricing models for valuing financial derivatives; namely, the constant elasticity of variance (CEV) model, the Bessel-K model, derived from the squared Bessel (SQB) process, and the unbounded Ornstein–Uhlenbeck (UOU) model, derived from the standard OU process. All three models are diffusion processes with linear drift and nonlinear diffusion coefficient functions. Specifically, the Bessel-K and UOU models are constructed based on a so-called diffusion canonical transformation methodology (Campolieti and Makarov, Int J Theor Appl Financ 10:1–38, 2007; Solvable Nonlinear Volatility Diffusion Models with Affine Drift, 2009; Math Finance 22:488–518, 2012). The models are calibrated to market prices of European options on the S&P500 index. It follows from the calibration analysis that the Bessel-K, UOU, and CEV models provide the best fit for pricing options that mature in 1 month, 3 months, and 1 year, respectively. The UOU model captures option data with a pronounced smile and hence it can be better calibrated to option data with short maturities. The CEV model provides a skewed local volatility and hence it works best for options with longer maturities. Moreover, we demonstrate that the CEV model is reasonably consistent through recalibration analysis on time series data in comparison with the Black–Scholes implied volatility.

1 Introduction

The Black–Scholes (BS) pricing formula for a standard call option is one of the most well-known formulae in mathematical finance. Regardless of its simplicity and reputation among scholars, it is a widely accepted fact that using it for pricing

A. S. Dahaj (✉)
University of Waterloo, Waterloo, ON, Canada
e-mail: a4soleim@uwaterloo.ca

G. Campolieti · R. N. Makarov
Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: gcampoli@wlu.ca

R. N. Makarov
e-mail: rmakarov@wlu.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_17

derivatives gives biased results due to its idealized assumptions such as log-returns being normally distributed and the volatility of the underlying stock price being constant. Evidences of extreme movements such as stock market crashes as well as volatility smiles and smirks greatly contradict the log-normality and constant volatility assumptions.

Local volatility models are one of the substitutes for the BS model. In local volatility models (known also as state-dependent volatility models), the volatility σ is itself a function of the current asset price S. This chapter studies three local volatility models namely the constant elasticity of variance (CEV), the Bessel-K, and the unbounded Ornstein–Uhlenbeck (UOU) [2–4]. The comparison of the models is done through the calibration process, which involves searching for parameter vectors that specify optimal dynamics of the models. The parameter values are found by minimizing the gap between market and model-generated call option prices [6]. To perform numerical tests we used Matlab with the optimization toolbox.

2 Data Extraction

Calibration of asset pricing models requires market option price data. Bloomberg Professional service which is a specialized software environment that allows its user to access and analyze real-time financial market data movements and place trades on the electronic trading platform [1], was used to extract historical option prices. Any financial instrument in Bloomberg has a specific Ticker by which it is known in the software. A Bloomberg terminal does not have any prespecified routine to automatically extract a time series of historical option prices in a required format. The appropriate syntax for locating an expired option is

TICKER: MM/YY OPTION TYPE <C> or <P> STRIKE <Yellow Key><GO>

So, as one can see, to find a specific expired option, the month and year of expiration, the type of option (Call or Put), and the strike should be provided. Upon loading the desired options, built-in Bloomberg functions can be used to get more information and also extract the historical prices of selected options. In this chapter, the historical option prices for SPDR S&P 500 ETF options (SPY) have been obtained and used for the calibration of the local volatility models. To extract and process expired option prices automatically, we wrote a code in Visual Basic (with the use of Bloomberg libraries) that allows us to connect to the Bloomberg Server and extract the data.

3 Asset Price Models

Consider a multiparameter diffusion model for the nonnegative asset price process $\{S_t\}_{t\geq 0}$ with the infinitesimal generator $(\mathcal{G} f)(S) \equiv \frac{1}{2}\sigma^2(S)S^2 f''(S)+rSf'(S)$, where $\sigma(S)$ is a nonlinear (local) volatility function and $r \geq 0$ is a risk-free interest rate.

For each model considered the discounted asset price is a martingale and hence we have no-arbitrage derivative pricing under each model.

The CEV diffusion has the power-type volatility function $\sigma(S) = \delta S^{\beta}$ with two parameters: $\delta > 0$ and $\beta < 0$. The point $S = \infty$ is hence a natural boundary for the process on $[0, \infty)$. For $\beta < -1/2$, the point $S = 0$ is a regular boundary, which we specify as killing, and for $-1/2 \le \beta < 0$ it is an exit boundary (see [7] for the boundary classification of diffusion processes). The dynamics of the CEV process relates to the non-central χ^2 probability distribution. The (risk-neutral) transition density has the known explicit representation:

$$p_S(t; S_0, S) = e^{-rt} \frac{(e^{-rt}S)^{-2\beta-\frac{3}{2}} S_0^{\frac{1}{2}}}{\delta^2|\beta|\tau_t} e^{-\frac{(e^{-rt}S)^{-2\beta}+S_0^{-2\beta}}{2\delta^2\beta^2\tau_t}} I_{\frac{1}{2|\beta|}}\left(\frac{(e^{-rt}S)^{-\beta}S_0^{-\beta}}{\delta^2\beta^2\tau_t}\right), \qquad (1)$$

for $S_0, S, t > 0$, $\tau_t = \frac{1}{2r\beta}\left(e^{2r\beta t} - 1\right)$ if $r \neq 0$ and $\tau_t = t$ if $r = 0$. This PDF involves $I_\alpha(z)$ which is the modified Bessel function of the first kind of order α and argument z.

The four-parameter Bessel K-family arises from a squared Bessel (SQB) process obeying the SDE $dX_t = \lambda_0 dt + \nu\sqrt{X_t}dW_t$, where we shall assume positive parameters $\mu \equiv \frac{2\gamma_0}{\nu^2} - 1$ and ν. The transition PDF p_S for the Bessel-K diffusion is related to the transition PDF p_X for the underlying SQB process as follows:

$$p_S(t; S_0, S) = \frac{\nu\sqrt{\mathrm{X}(S)}}{\sigma(S)S} \frac{u_\rho\left(\mathrm{X}(S)\right)}{u_\rho\left(\mathrm{X}(S_0)\right)} e^{-\rho t} p_X(t; \mathrm{X}(S_0), \mathrm{X}(S)). \qquad (2)$$

Here, $\mathrm{X} \equiv \mathrm{F}^{-1}$ is the unique inverse of the map $\mathrm{F}(x) = c\frac{I_\mu(2\sqrt{2(\rho+r)x}/\nu)}{K_\mu(2\sqrt{2\rho x}/\nu)}$, and $u_\rho(x) = x^{-\mu/2}K_\mu\left(2\sqrt{2\rho x}/\nu\right)$ with $\rho > 0$ is the so-called generating function. The volatility function is expressed in term of the modified Bessel functions I and K (see [3] and [4] for more details); it has the following asymptotic properties:

$$\sigma(S) \propto S^{-\frac{1}{2\mu}}, \text{ as } S \to 0, \text{ and } \sigma(S) \to \text{Constant, as } S \to \infty.$$

By choosing the Ornstein-Uhlenbeck (OU) diffusion (solving the SDE $dX_t = (\lambda_0 - \lambda_1 X_t)dt + \nu_0 dW_t$) as an underlying diffusion, we obtain the (OU) families of diffusions with four parameters. For details, see [3] and [4]. The volatility of the UOU model has the following asymptotic behavior:

$$\sigma(S) \propto \sqrt{|\ln S|}, \text{ as } S \to 0, \text{ and } \sigma(S) \propto \sqrt{\ln S}, \text{ as } S \to \infty.$$

4 Calibration of Asset Price Models

Model calibration consists of finding an optimal parameter vector, denoted by ξ, that specifies the asset price model such as the CEV, Bessel-K, or the UOU model. To measure the distance between market and model prices, we use the mean square error

Table 1 Comparing models calibrated based on single maturity SPY market option data on "19/10/2007." The residuals ε are reported

Maturity date	CEV	Bessel-K	UOU
11/17/2007 (nearest to 1 month)	0.411	0.079	0.126
3/22/2008 (nearest to 6 months)	0.208	0.185	0.064
9/20/2008 (nearest to 1 year)	0.169	0.377	0.338

(MSE) in a loss function to calibrate the vector of model parameters ξ. Consider a standard call option with strike K_i and time to maturity T_i having an observed market price V_i^{mkt}, while its value under the specified model is $V(K_i, T_i, \xi)$, where $i = 1, 2, \ldots, N$. Here, N is total number of data contracts used in the calibration. The loss function is then

$$F(\xi) = \sum_{i=1}^{N} (V(K_i, T_i, \xi) - V_i^{mkt})^2. \tag{3}$$

Consequently, the vector $\hat{\xi}$ of calibrated model parameters minimizes the function $F(\xi)$,

$$\hat{\xi} = \arg\min_{\xi}\{F(\xi)\}\,. \tag{4}$$

The residual $F(\hat{\xi})$ is then converted to $\varepsilon = \sqrt{F(\hat{\xi})/N}$ and the later is used in comparing the calibration results for different underlying assets (or indices), dates, and models.

The parameter vector of the CEV model to be calibrated is $\xi = (\beta, \sigma_0)$, where $\sigma_0 \equiv \delta S_0^{\beta}$ denotes the instantaneous volatility at the spot S_0. For the Bessel-K model, we have a four-parameter vector $\xi = (\mu, \nu, \rho, c)$. After performing test runs on the Bessel-K model, we found that ν is a redundant parameter, so we simply set $\nu = 2$. The four-parameter vector for the UOU model is $\xi = (\rho, \upsilon, \kappa, c)$. Likewise, after performing several test runs on the UOU model, we found that κ is a redundant parameter so we set $\kappa = 1$.

4.1 Comparison of the Models

It makes more sense that each model can capture a specific type of skewness of local volatility. The local volatility of the Bessel-K is more skewed and that of the UOU model has a pronounced smile shape. Similarly, the BS implied volatility for short maturities is more smile-like and for long maturities it is more skewed. Table 1 presented below shows that the Bessel-K model fits the best for the nearest to 1 month maturity options, the UOU model fits the best for nearest to 6 months maturity, and the CEV model fits slightly better than the Bessel-K and UOU models for nearest to 1 year maturity.

Table 2 Calibration results of the constant elasticity of variance (CEV) model of the underlying SPY based on 06/1/2006 data

Date	S_0	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\delta}$	ε
06/1/2006	128.46	−0.0918	0.1137	0.1776	0.8229

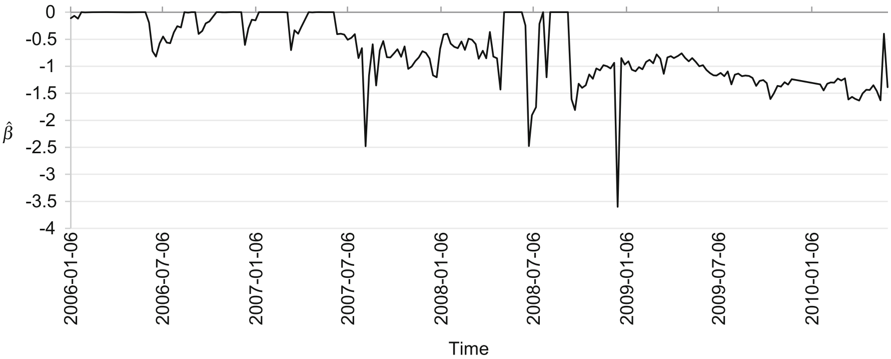


Fig. 1 SPY weekly recalibration results: $\hat{\beta}$ time series

4.2 Recalibration of the CEV Model

The recalibration essentially means doing the calibration on time series of arrays of option data. The starting points for each element of the time series of option data is the calibrated parameter vector from the preceding element of the time series:

$$\xi^s_{n+1} = \hat{\xi}_n, \quad n = 1, 2, 3, \ldots \tag{5}$$

where ξ^s_{n+1} is the starting parameter vector in the $(n+1)$th iteration, and $\hat{\xi}_n$ is the calibrated parameter vector in the nth iteration. We used weekly call option data for the underlying SPY starting at “06/1/2006,” first Friday of 2006, to “6/4/2010.”

The starting parameter values are derived by applying the CEV calibration routine for the first element (i.e., first date) of the time series and the results of the calibration can be found in Table 2.

The recalibration results for $\hat{\beta}$ can be seen in Fig. 1, which shows that volatility steepness parameter $\hat{\beta}$ in the CEV model through the time series varies considerably.

The recalibration results for $\hat{\sigma}_0$ can be seen in Fig. 2. Since $\hat{\sigma}_0$ is the calibrated local volatility at the money, we compare $\hat{\sigma}_0$ with the BS implied volatility at the money with the maturity nearest to 1 year. We observe that the CEV model calibration captures $\hat{\sigma}_0$ well enough, as the local volatility of the CEV model at S_0. The BS implied volatility also follows the same trend.

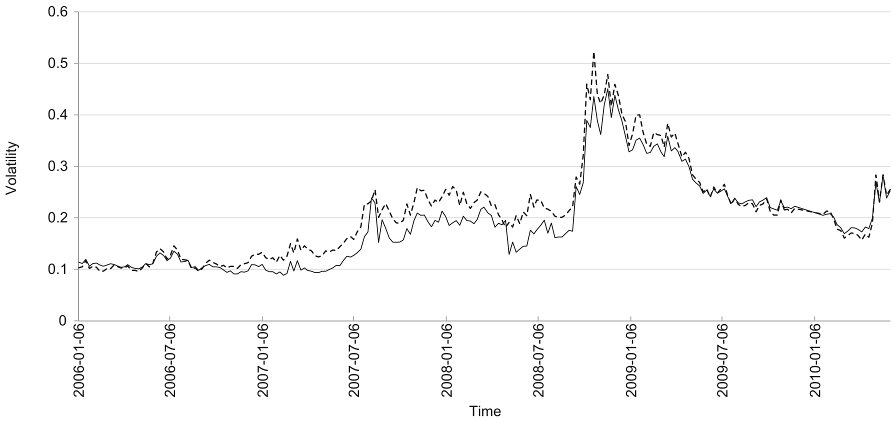


Fig. 2 SPY weekly recalibration results: $\hat{\sigma}_0$ time series (*solid line*). The BS implied volatility at the money with maturity 1 year is given by a *dashed line*

References

1. Bloomberg Professional: http://www.bloomberg.com/professional/
2. Campolieti, G., Makarov, R.: Pricing path-dependent options on state dependent volatility models with a Bessel bridge. Int. J. Theor. Appl. Financ. **10,** 1–38 (2007)
3. Campolieti, G., Makarov, R.: Solvable nonlinear volatility diffusion models with affine drift. Arxiv preprint. arXiv:0907.2926 (2009)
4. Campolieti, G., Makarov, R.: On properties of analytically solvable families of local volatility diffusion models. Math. Financ. **22,** 488–518 (2012)
5. Cox, J.C.: Notes on Option Pricing I: Constant Elasticity of Variance Diffusions. Unpublished manuscript, Stanford University, Graduate School of Business (1975)
6. Fusai, G., Roncoroni, A.: Implementing models in quantitative finance: methods and cases. Springer Finance. Springer, Berlin (2008)
7. Karlin, S., Taylor, H. M.: A second course in stochastic processes. Academic Press, Inc., New York-London (1981)

An Application of the Double Skorokhod Formula

Cristina Canepa and Traian A. Pirvu

Abstract This chapter considers the problem of borrowing and lending federal funds by a bank. The goal of the bank is to find the optimal borrowing/lending transaction policy while maintaining the reserve requirements. Within the model of [3] and [6] we describe the optimal net transaction amount using the Skorokhod formula developed in [8]. This formula provides a fast way of computing the optimal net transaction amount.

1 Introduction

Assume an economy with only one bank and the Federal Reserve Bank; the bank's task is to derive an optimal transaction amount to minimize the cost of buying and selling funds from the Federal Reserve Bank, while meeting the reserve requirements. The reserve requirements are determined by the demand deposit flow. The bank meets its reserve requirements if the excess reserve process (i.e. the difference between the reserves and the required reserves) remains positive. We assume that during the business day, the bank can increase/decrease its level of federal funds through direct transactions, which involve transaction costs. The net deposit flow is exogenously specified and is modelled as a Brownian motion with drift. We assume that the bank is a price-taker in the federal funds market and can obtain sufficient credit from the Federal Reserve Bank. The bank's objective is to choose the federal fund purchases and sales in order to minimize the cost function. The optimal net transaction amount is expressed by the Skorokhod formula developed in [4].

C. Canepa (✉)
Faculty of Mathematics, University of Bucharest,
Bulevardul Regina Elisabeta 4-12, Bucharest, Romania
e-mail: elenacristina2@gmail.com

T. A. Pirvu
Math and Stat Department, McMaster University,
1280 Main St W, Hamilton, ON, Canada.
e-mail: tpirvu@math.mcmaster.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_18

2 The Model

This section describes the model employed. The inputs are described together with the mathematical framework.

There are two levels of uncertainty, corresponding to the macro level and the micro level. The *economy* is characterized by **A**, a random variable that describes the asset sizes, following the distribution f_A. A *macroeconomical policy* is characterized by a set of strictly positive parameters (λ, q), where:

1. $\lambda > 0$ is the target interest rate imposed by the Federal Reserve Bank; it is used as a discount rate in the model.
2. $q \in [0, 1]$ is the fraction of the deposits that are required to be kept as reserves.

At the *micro level*, one bank is characterized by:

1. An asset size A, which is a realization of the random variable **A**.
2. An exogeneously given demand deposit process $(D_t)_{t\geq 0}$.
3. A required reserve process $(R_t)_{t\geq 0}$, where $R_t = qD_t$ at every time $t \geq 0$.
4. An excess reserve process $(X_t)_{t\geq 0}$, where $X_t = (1-q)D_t$ for every time t.
5. A net purchase amount process $(W_t)_{t\geq 0}$, which is the result of an optimal control problem.

We are interested in a model that connects the excess reserve process to the optimal transaction amount that a bank proposes to buy or sell.

2.1 Mathematical Framework

In this subsection the mathematica setup is described. The deposit process is driven by a Brownian motion with drift.

Let (Ω, F, P) be a probability space on which we consider the asset size random variable **A** and the continuous excess reserves process $(X_t)_{t\geq 0}$ (which are not necessarily independent). The probability measure P is the real world probability measure. We consider $\mathbf{F} = (F_t)_{t\geq 0}$ to be the completion of the augmented filtration generated by X (so that $(F_t)_{t\geq 0}$ satisfies the usual conditions). Therefore, the bank observes nothing except the sample path of X. We assume $X_0 = x \geq 0$ with probability 1. A larger filtration is given by the regulator's filtration, $(G_t)_{t\geq 0}$, where G_t is the completion of the σ-algebra generated by F_t and **A**. A standard, one-dimensional, Brownian motion is a continuous, adapted process $B = (B_t, \mathrm{F}_t, 0 \leq t \leq \infty)$, with the property that $B_t - B_s$ is independent of F_s and normally distributed of mean 0 and variance $t - s$. We consider that the deposit process that corresponds to one bank is exogeneously given and it follows a Brownian motion with drift:

$$dD_t = \tilde{\mu} dt + \tilde{\sigma} dB_t. \tag{1}$$

The exogeneously given *market* is characterized by: $(\alpha, \beta, h, \tilde{\mu}, \tilde{\sigma})$, where:

1. (α, β, h) are deterministic functions of the asset size A that express the transactions costs for buying, selling and holding funds, respectively.

2. $(\tilde{\mu}, \tilde{\sigma})$ are deterministic functions of the asset size A that express the Brownian motion parameters driving the deposit processes.

Correspondingly, the excess reserve process follows a Brownian motion with drift with modified parameters:

$$dX_t = \mu dt + \sigma dB_t. \tag{2}$$

The parameters μ, σ are also deterministic functions of the asset size A[1]. It is considered that banks/financial companies incur three types of costs in managing their excess reserve positions:

1. A proportional transaction cost α of buying funds.
2. A proportional transaction cost β of selling fed funds.[2]
3. A continuous holding cost, incurred at the rate hX_t.

It is assumed that $\alpha + \beta > 0$ for, otherwise, the bank would be allowed to have profit without taking any risk, thus giving rise to arbitrages. In addition, we also assume that $\beta < h/\lambda$ for otherwise it is never optimal to sell.

3 Problem Formulation

In this section we formulate the objective of this chapter. The policies are defined and the goal is to minimize a cost function over the set of feasible policies.

Let us start with the following formal definition.

Definition 1 A policy is defined as a pair of processes L and U such that

$$L, U \quad \text{are} \quad \mathbf{F} - \text{adapted, right-continuous, increasing and positive.} \tag{3}$$

In the context of the federal funds market, L_t and U_t are the cumulative increases (federal funds purchases) and decreases (federal funds sales) that the bank undertakes up to time t, in order to satisfy the reserve requirements and to maximize its profit.

Definition 2 A controlled process associated to the policy (L, U) is a process $Z = X + L - U$.

In our model for the federal funds market, Z_t is the amount of excess funds in the bank's reserve account.

Definition 3 The policy (L, U) is said to be feasible if

$$L_{0-} = U_{0-} = 0, \tag{4}$$

[1] Since $X_t = (1 - q)D_t$ we can express the parameters μ, σ in terms of $\tilde{\mu}, \tilde{\sigma}$ and q

[2] The proportional adjustment costs, α and β, are due to spreads between bid and ask prices, brokerage fees, the lack of availability of a transaction partner and other service charges which vary with the volume of the transaction, as in [2].

$$P_x\{Z_t \geq 0, \forall t\} = 1, \forall x \geq 0, \tag{5}$$

$$E_x\left[\int_0^\infty e^{-\lambda t} dL\right] < \infty, \forall x \geq 0, \tag{6}$$

and

$$E_x\left[\int_0^\infty e^{-\lambda t} dU\right] < \infty, \forall x \geq 0. \tag{7}$$

We denote by $\tilde{S}(x)$ the set of all feasible policies associated with the continuous process X that starts at x.

Definition 4 The *cost function* associated to the feasible policy (L, U) is

$$k_{L,U}(x) \equiv E_x\left[\int_0^\infty e^{-\lambda t}(hZ_t dt + \alpha dL + \beta dU)\right], \qquad x \geq 0. \tag{8}$$

Definition 5 The control $(\hat{L}, \hat{U})$ is said to be *optimal* if $k_{\hat{L},\hat{U}}(x)$ is minimal among the cost functions $k_{L,U}(x)$ associated with feasible policies (L, U), for each $x \geq 0$.

As in [3], by restricting to barrier type policies and by assuming a given upper barrier $b > 0$, the bank's problem is to find the transaction amount that keeps the controlled excess reserve process Z between 0 and b. Therefore, the problem connects to the double Skorokhod map.

3.1 Double Skorokhod Map

In this section we give a short background on the double Skorokhod map.

Let $D[0,\infty)$ be the space of positive, right-continuous functions with left limits mapping $[0,\infty)$ into R.

Definition 6

The double Skorokhod map $\Gamma_{0,b}$ is the mapping of $D[0,\infty)$ into itself such that for every $\psi \in D[0,\infty)$, $\Gamma_{0,b}(\psi)$ takes values in $[0,b]$ and has the decomposition

$$\Gamma_{0,\mathrm{b}}(\psi) = \psi + \eta_l - \eta_\mathrm{u}.$$

Here η_l, η_u are nondecreasing functions in $D[0,\infty)$ so that the triple $(\Gamma_{0,b}(\psi), \eta_l, \eta_u)$ satisfies the complementary conditions

$$\int_0^\infty I_{\Gamma_{0,b}(\psi)(s)>0} d\eta_l(s) = 0, \qquad \int_0^\infty I_{\Gamma_{0,b}(\psi)(s)<b} d\eta_u(s) = 0. \tag{9}$$

An explicit formula for the double Skorokhod map $\Gamma_{0,b}$ on the space $D[0,\infty)$ was recently obtained in [4],

$$\Gamma_{0,b}(\psi)(t) = \psi(t) - [(\psi(0)-b)^+ \wedge \inf_{u\in[0,t]} \psi(u)] \vee \sup_{s\in[0,t]} [(\psi(s)-b) \wedge \inf_{u\in[s,t]} \psi(u)] \tag{10}$$

3.2 The Optimal Net Transaction Amount

This subsection contains the main result of the chapter. It provides the optimal net transaction amount by means of the Double Skorokhod Formula.

According to [1] and [3] (Chapter 'Solving the Linear Transportation Problem by Modified Vogel Method'), the optimal strategy turns out to be a barrier strategy, i.e.

1. $(\hat{L}, \hat{U})$ continuous on $(0, \infty)$, increasing, $\hat{L}_{0-} = \hat{U}_{0-} = 0$,
2. $Z_t \equiv X_t + \hat{L}_t - \hat{U}_t \geq 0, \forall t \geq 0$,
3. $\int_0^t I_{Z_t > 0} d\hat{L}_t = 0, \int_0^t I_{Z_t < b} d\hat{U}_t = 0$.

The upper barrier b solves the equation

$$g(-b) = g(0)\frac{h + \lambda\alpha}{h - \lambda\beta}. \tag{11}$$

Here the function g is defined by

$$g(x) \equiv \gamma_1 e^{\gamma_2 x} + \gamma_2 e^{-\gamma_1 x}, \tag{12}$$

and γ_1, γ_2 are the roots of

$$\sigma^2\gamma^2/2 + \mu\gamma - \lambda = 0.$$

The next theorem is our main result.

Theorem 1 *The bank's optimal net transaction amount* $\hat{L} - \hat{U}$ *is given by*

$$\hat{L}_t - \hat{U}_t = -[(X_0 - b)^+ \wedge \inf_{u \in [0,t]} X_u] \vee \sup_{s \in [0,t]} [(X_s - b) \wedge \inf_{u \in [s,t]} X_u]. \tag{13}$$

Proof Double Skorokhod Formula of [4] gives (13). □

References

1. Canepa, C.: Numerical simulation of defaults in large banking systems. Ph.D. thesis, Carnegie Mellon University (2012)
2. Chen, A., Mazumdar, S.: An instantaneous control model of bank reserves and federal funds management. J. Bank. Financ. **16,** 1073–1095 (1992)
3. Harrison, M.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
4. Kruk, L., Lehoczy, J., Ramanan, K., Shreve, S.: Double skorokhod map and reneging real-time queues. In: Ethier, S., Feng, J., Stockbridge, R. (eds.) Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz. vol. 4, pp. 169–193. Institute of Mathematical Statistics Collections (2008) doi:10.1214/074921708000000372

Multitaper Smoothed Minimum Statistics Noise Power Estimation

Ricardo Castellanos, Nurgun Erdol and Hanqi Zhuang

Abstract Speech communication devices and digital hearing aids must perform in the presence of high levels of ambient noise. Speech enhancement is a denoising process where Wiener-like filters are developed that require the estimation of the background noise spectrum from an additive combination of speech and noise. To follow statistical variations over time, the processes must be performed over short and overlapping frames of data resulting in time varying filters and spectra. We propose a novel algorithm to track the noise power of each frequency bin as it evolves over time. The proposed method uses an adaptation of the multitaper autoregressive spectral estimate. The resulting spectral components are smooth, low bias, and low variance and show superior tracking of the time-variation of the spectra.

1 Introduction

Noise power estimation is an essential task in the speech-enhancement process. Speech enhancement has become increasingly important as speech-processing devices, such as mobile phones and hearing aids, have risen in popularity, and users expect them to work everywhere under many different conditions where acoustic disturbances may degrade the quality of audio leading to user discomfort.

The performance of the speech-enhancement system depends on the accurate estimation of the noise spectrum that shows the distribution of its power over frequency. When the system overestimates the noise power, speech components are distorted

R. Castellanos (✉) · N. Erdol · H. Zhuang
Florida Atlantic University, 777 Glades Rd, Boca Raton, FL 33431, USA
e-mail: rcastel5@fau.edu

N. Erdol
e-mail: erdol@fau.edu

H. Zhuang
e-mail: zhuang@fau.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_19

degrading the intelligibility of the speech and often introducing musical noise which is considered to be annoying. Underestimating the noise power that falls short of the desired noise-reduction goal, may impair intelligibility and introduce stress.

Speech enhancement systems using a single-microphone have over three decades of research. Some methods are based on voice activity detectors (VAD) where the noise estimate is updated during speech absence moments, that is, the noise power estimate is given by the mean of the noisy signal over speechless segments. This approach is highly dependent on the speech power so its performance and reliability can be seriously reduced at low-input signal to noise ratio (SNR). Other methods based on histograms in the power spectral domain require a lot of computational power and memory resources and their performance is poor at low SNR conditions.

A method based on minimum statistics for noise estimation obtains the noise estimate tracking the minima of a smoothed power spectral estimate of the noisy signal. However, it is required to multiply the estimate by a factor to compensate for bias, and the variance of the estimate is high and very sensitive to outliers.

A more recent method called improved minima controlled recursive averaging (IMCRA) combines the minimum tracking approach with recursive averaging of the past spectral estimates using a smoothing parameter that is adapted frame by frame by the probability of speech presence in different frequency subbands. Even though the algorithm performance is good for enhancing speech, the recursive averaging falls short of tracking the spectral minima.

In this chapter, we propose a method to track the noise power by adapting the multitaper autoregressive (MTAR) spectral estimation algorithm. The adaptation fits an autoregressive (AR) spectrum on the fixed frequency component of a time-frequency data obtained by a spectrogram. The estimate has low bias and variance and tracks changes in the noise spectrum more closely than its predecessor methods. The method also eliminates the need to estimate the probability of speech presence, which is not highly reliable.

This chapter is organized as follows: Sects. 2 and 3 present review of speech enhancement and MTAR spectral estimation. The adaptation of the MTAR to smoothing the time evolution of the spectral components is presented in Sect. 4. Experimental results and conclusions are given in Sect. 5.

2 Speech Enhancement

Let $x[n]$ and $d[n]$ represent samples, respectively, of the speech and noise signals at the nth sampling point. Prevalent and reasonable assumptions are that speech and noise combine additively to form the noisy speech signal and are uncorrelated. Speech signals are always and noise is frequently nonstationary which necessitates that they be analyzed and processed in short time frames over which stationarity assumptions are more or less true. The observed noisy speech data are organized in overlapping frames of the average length of a phoneme and are analyzed by the short-time Fourier transform (STFT) given by

$$Y\,[k,l] = \sum_{n=0}^{N-1} y\,[n + lM]\,h\,[n]\,e^{-j\frac{2\pi}{N}nk} \tag{1}$$

Here k corresponds to the frequency bin index, l is the frame number, $h[n]$ is the analysis window of size N, and M is the framing step. The short time spectral amplitude estimate of the clean speech is given by the action of the gain function $G\,[k,l]$ on the Fourier transform of the noisy speech as given by

$$\hat{X}[k,l] = G[k,l]Y[k,l] \tag{2}$$

The gain function produces an estimate of the spectral components of the enhanced speech $\hat{X}[k,l]$ from the given noisy spectral components $Y[k,l]$. The success of the enhancement process is critically dependent on the estimation of the noise power spectrum.

A commonly used approach is to compute or update noisy statistics over non-speech segments that are detected by voice activity or a speech pause detectors. Impediments to success are decline in detection reliability under conditions of low segmental SNR and the low number of segments or frames that are completely speech free.

The minimum statistics approach is based on the observation that even during speech activity the power spectral density of the noisy speech repeatedly decays to values that are comparable to the noise power level [5]. Therefore, by tracking the minimum of the noisy speech, the system can derive an estimate of the noise power within a finite window. The method offered in [5] is sensitive to outliers and has variance. Cohen et al [1, 2] proposed *the minima controlled recursive averaging* method that averages past spectral values using a recursive smoothing filter. The filter coefficients are adjusted by the probability of speech presence in the segment.

Our approach derives the recursive smoothing filter coefficients using an adaptation of the MTAR [3] spectral estimator on the spectral components as they evolve through time. The result is a low-bias and low-variance smoothing of the spectrum without the cumbersome computation of speech probabilities.

3 Multitaper Autoregressive Spectral Estimate

The multitaper autoregressive (MTAR) Spectral Estimate

$$\hat{S}_{MTAR}(\omega) = G\left|1 - \sum_{k=1}^{p} a_k e^{-j\omega k}\right|^{-2} \tag{3}$$

uses the multitaper autocorrelation (MTAC) estimates of lags up to order p to estimate the coefficients $\{G, a_1, \ldots, a_k\}$ of (3). The filter coefficients are found by solving the Yule–Walker equations

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \tag{4}$$

where $\mathbf{R}[j,k] = \hat{r}[|j-k|]$ and $\mathbf{r}[k] = \hat{r}[k]$ for $\{j,k = 1,2,\ldots,p\}$ are the MTAC estimated from the data $\{x[n],\ n = 0,1,\ldots,N-1\}$. The MTAC are derived from the MT spectral estimate [6,1] which is the average of direct estimates $|X_k(\omega)|^2$

$$\widehat{S}(\omega) = \frac{1}{K}\sum_{k=0}^{K-1}|X_k(\omega)|^2 \tag{5}$$

where $X_k(\omega) = \sum_{n=0}^{N-1} x[n]v_k[n]e^{-j\omega n}$ is the discrete-time Fourier transform (DTFT) of data sequence multiplied by the taper $v_k[n]$. The tapers are typically vectors of a complete basis in R^N, the space of N-length vectors and $K << N$. The original tapers used by Thomson are the discrete prolate spheroidal or Slepian sequences (DPSS). Minimum-bias and their approximate sine tapers have also been used successfully [3].

The MTAC estimate $\widehat{r}[n]$ is the inverse transform (IDTFT) of the MTSE $\widehat{S}(\omega)$ (5) [3, 4] and is given by

$$\widehat{r}[n] = \sum_{m=0}^{N-1} x[m]x[m-n]\alpha_{K,N}[m,n] \tag{6}$$

Where $\alpha_{K,N}[m,n] = \frac{1}{K}\sum_{k=0}^{K-1} v_k[m]v_k[m-n]$. The expected value of the MTAC (6) yields

$$r[n] = r_x^s[n]w[n] \tag{7}$$

where

$$w[n] = \sum_{m=0}^{N-1}\alpha_{K,N}[m,n] \tag{8}$$

is the lag window and $r_x^s[n] = E\{x[m]x[m-n]\}$ is the autocorrelation of the stationary process. The lag window is the average of deterministic autocorrelations and therefore it is symmetric and has a nonnegative Fourier spectrum. The properties of the estimator are analyzed in a companion paper [4] and are shown to be asymptotically consistent for lags $n < N/K$. K is adjusted for the desired spectral concentration bandwidth of the DPSS tapers. It is a small number, typically 3–4 for a data length of 100. MTAR spectral estimation uses the MTAC estimates of lags determined by the system order

4 Smoothing of the Time Evolutionary Spectra by the Adapted MTAR Algorithm

The contribution of this work is related to the way the spectral estimator is evaluated from the spectrogram of the noisy speech. We take advantage of the low bias and variance of the MTAR estimate and its capability to represent a smoothed version of the periodogram by taking horizontal cross sections of the spectrogram that correspond to the periodogram of a specific frequency bin as it changes over frames. With reference to (1), let $S_l\,[k] = |Y\,[k,l]|^2$ be the periodogram estimate of frame as a function of the frequency bin index k. Alternatively, we can represent the spectrogram by $\tilde{S}_k\,[l] = |Y\,[k,l]|^2$ for a fixed frequency bin k as a function of the frame number or corresponding time. It is shown in the plot directly.

Even though it is a function of time, $\tilde{S}_k\,[l]\,, l = 0, 1, \ldots, L-1$ has all the properties of a spectrum except the symmetric component for negative values of l. In a step we call spectrizing, we create the symmetric sequence $\phi_k\,[l] = \phi_k\,[-l] = \tilde{S}_k\,[l]$ that is a spectrum. Its inverse discrete Fourier transform (DFT) is an autocorrelation sequence $r_k\,[l]\,, l = 0, \pm 1, \ldots \pm L - 1$. We multiply by the AC sequence with the multitaper lag window (8) to obtain

$$\rho_k\,[l] = w\,[l]\,r_k\,[l] \tag{9}$$

and use $\rho_k\,[l]\,, l = 0, 1, \ldots, p$ in the Yule–Walker (YW) equations (4). The solution to the YW equations yields the coefficients of an all-pole filter and a gain constant. The spectral estimate is then obtained in accordance with (3).

5 Experiments, Results and Conclusions

In order to test the effectiveness of the MTAR estimates, we substituted the estimate given by the IMCRA algorithm by the MTAR estimate so it was possible to analyze the performance of the estimator in comparison with the IMCRA method. Additionally, MTAR estimates were evaluated under different configurations:

- MTARfull: Using the whole set of data points (full frame size) when applying the “spectrize” approach so the MTAR estimate corresponds to the whole sequence for every single frequency bin in the spectrogram using a filter order of 50 and 4 tapers.
- MTARwin: Using a windowed approach with overlapping frames of 15 samples with 14 overlapping samples using a filter order of 3 and 4 tapers.
- Using four different types of tapers: Sine tapers, Slepian sequences, Slepian sequences with twice the number of data samples ($2N$) but truncated to N samples, and Slepian sequence combined with Sine tapers.

The input signals for noise belong to the Noisex92 database which comprises White Gaussian Noise (GWN), Car noise, and F16 cockpit noise among others. The input signals for speech belong to the TIMIT database. Speech was degraded with noise at different SNR values in the range of −5 to 5

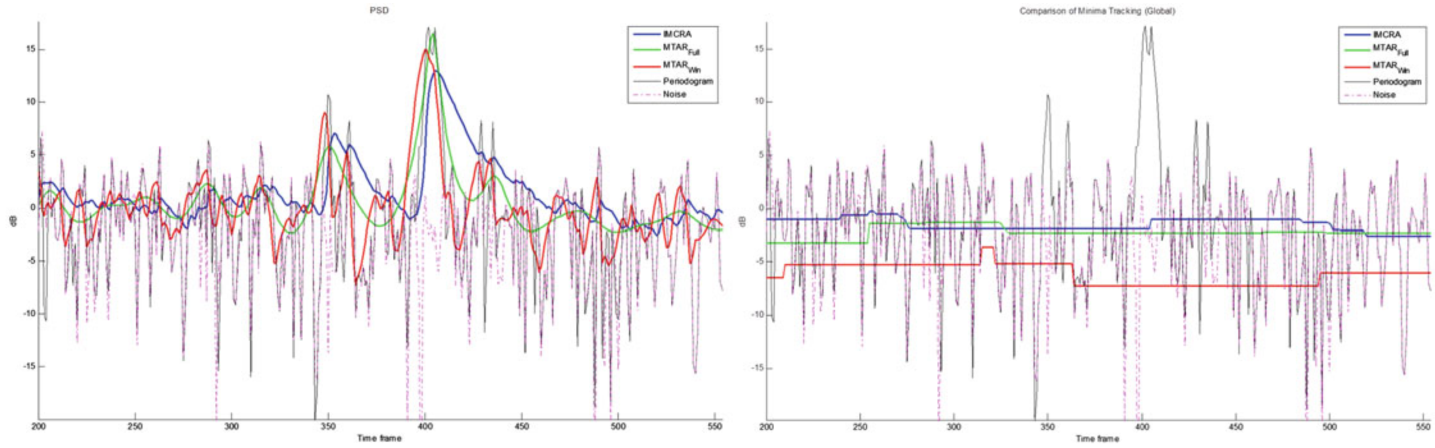


Fig. 1 (*Left*) Power Spectral Density. (*Right*) Comparison of tracking

The capability of the MTAR estimator to track the changes better in the smoothed spectrum compared to the IMCRA estimator has been illustrated. In both cases, MTARfull and MTARwin were able to follow the changes of the periodogram with higher accuracy being capable to reach higher and lower local values.

The improvement of the MTAR estimates having lower segmental error for all the different tapers was tested. The combination of Slepian sequences with Sine tapers showed better results for both MTARfull and MTARwin estimates.

References

1. Cohen, I.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. Signal Process. Lett. **9**(1), 12–15 (2002)
2. Cohen, I.: Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. Trans. Speech Audio Process. **11**(5), 466–475 (2003)
3. Erdol, N., Gunes, T.: Multitaper covariance estimation and spectral denoising, Thirty-Ninth Asilomar Conf. Signals Syst. Comput. **1,** 1145–1147 (2005)
4. Erdol, N.: The Multitaper Covariance and Autoregressive Spectral Estimates, Submitted to the Proceedings of the AMMCS (2013)
5. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. Trans. Speech Audio Process. **9,** 504–512 (2003)
6. Thomson, D.J.: Spectrum estimation and harmonic analysis. Proc. IEEE **70**(9), 1055–1092 (1982)
7. Thomson, D.J.: Some Comments on Multitaper Estimates of Autocorrelation, Proceedings Statistical Signal Processing Workshop(SSP), pp. 656–659 (2012)

Design Considerations for Thermal Management of Electronics Enclosures

Rachele Cocks, David Clendenen and Ludovic Chretien

Abstract This study focuses on the design and optimization of an electronics enclosure intended for operation in an outdoor commercial heating, ventilation, and air conditioning, HVAC, application. In particular the design was optimized for a high ambient environment without the aid of forced air cooling. As electronically controlled motor drive systems are increasing in use, designs need to operate in new challenging environments, reach higher power density, and enable higher levels of system integration. Computational fluid dynamics, CFD, was used for design, analysis, and optimization and correlated with test data. A design of experiments, DOE, was used to evaluate the sensitivity of the final design to the operating environment. A constrained optimization was performed to determine the optimal fin spacing, height, angle, and thickness of the enclosure geometry for thermal dissipation of the heat from the power electronics. Various fin topologies were also analyzed to evaluate the impact of increased surface area and enhanced thermal mixing effects. After a thorough review of the design space, general design recommendations are made and an optimized design reviewed.

1 Introduction

As electronically controlled motor drives are increasing in use, designs need to operate in new challenging environments, reach higher power density, and enable higher levels of system integration.

This chapter focuses on the design and optimization of an electronic enclosure intended for operation in an outdoor commercial heating, ventilation, and air conditioning application. The first section of the chapter will explore the characteristics

R. Cocks (✉) · D. Clendenen · L. Chretien
Regal Beloit Corporation, Fort Wayne, IN, USA
e-mail: Rachele.Cocks@RegalBeloit.com

D. Clendenen
e-mail: David.Clendenen@RegalBeloit.com

L. Chretien
e-mail: Ludovic.Chretien@RegalBeloit.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_20

of the electronic drive components as well as the machine control strategy and the impact on power losses and therefore thermal performance. The later section will present the results of the enclosure optimization through design of experiment (DOE) and use of computational fluid dynamics (CFD).

2 Design Space and Application

A motor for heating ventilation, and air conditioning (HVAC) systems is considered in this study. This particular motor is to be used in large condenser fan units that comprise a compressor, evaporation coils, and the fan motor. These units can be arranged in single fan or multiple fan configurations. The operating environment can reach up to 60 °C when the units are placed on rooftops, and depending on the climate of the region of installation.

Also to be considered is the ingress protection (IP) rating of the motor, as well as its construction. Some motors can be totally enclosed fan cooled (TEFC), with the fan being located either inside or outside the enclosure.

In addition to the application environmental requirements, the motor construction topology as well as the ingress protection rating significantly impacts the system thermal characteristics. For our study, the motor is TEFC, with an internal fan providing cooling for machine and electronics.

3 Thermal Management of Electronics

While focusing on the thermal management of the electronic drive, it is necessary to evaluate the thermal characteristics of the electric machine as the motor assembly combines machine and electronics in an integrated package.

3.1 Drive Overview

The typical heat producing and heat sensitive components of an electronic drive are represented in Fig. 1. Amongst these components are the inrush NTC, used to limit the amount of current drawn from the AC line when the DC capacitors are completely discharged, the inductance part of the input filter used to mitigate electromagnetic interferences (EMI), and the power devices: low speed diodes for the AC/DC section of the drive, and IGBT/Diodes used for the DC/AC conversion to drive the machine windings.

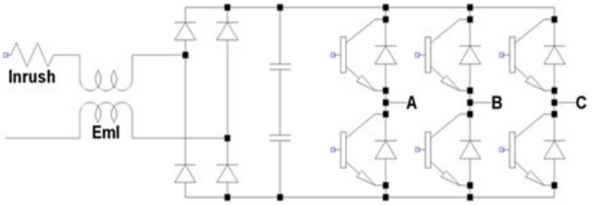


Fig. 1 Principal heat sources and heat sensitive components

3.2 Passive Components

The inrush NTC, unless taken out of the circuit by some mechanical means (relay) or electronic device (solid state switch), can operate at temperatures reaching up to 125 °C, and significantly affect the internal ambient of the electronic drive.

The electronics internal ambient can have a critical impact on performance and long term reliability. For example, the operating temperature of the main electrolytic capacitors determines how long they will last in the application. From a performance standpoint, the operating temperature of the EMI choke needs to be kept low enough to avoid saturation and a decrease in its filtering capability.

3.3 Power Devices and Motor Control Strategy

Careful selection of the electronic controller power devices is necessary to ensure adequate thermal performance. Thermal performance of these devices is related to the losses they have to dissipate. In the case of a passive diode bridge rectifier, the losses equation is described in Eq. 1, where V_f is the forward voltage drop of the diode and I_f is the current flowing through the device. The bridge losses can be affected by the technology of the diodes used (different V_f), as well as the line impedance feeding the system. Indeed, different line impedance can affect the system power factor and cause it to vary.

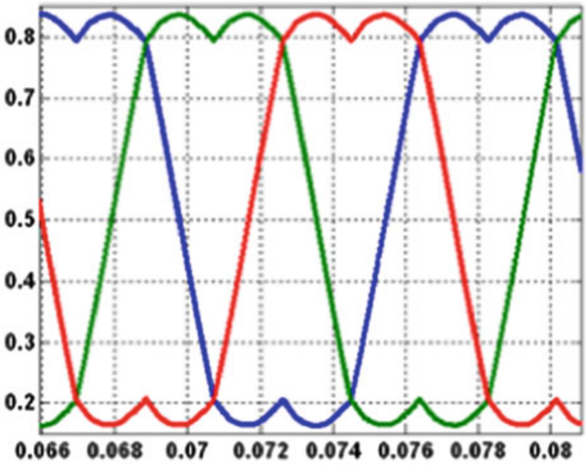
$$P = \frac{1}{T}\int V_f(t)\cdot I_f(t)dt \tag{1}$$

The losses for the IGBT/Diode inverter devices can be divided in two categories: conduction losses and switching losses. The IGBT conduction losses are described in Eq. 2, and depend on the collector to emitter saturation voltage of the component, as well as the current flowing through it. The conduction losses should be evaluated at machine peak operating current.

$$P = \frac{1}{T}\int V_{cesat}(t)\cdot I_f(t)dt \tag{2}$$

The IGBT switching losses are described in Eq. 3. These losses occur when the current flow changes from a lower switch to an upper switch (and inversely) within

Fig. 2 Standard SVPWM(space vector pulse width modulation) switching pattern



an inverter leg. E_{on} and E_{off} represent the energy seen in the IGBT at turn on and turn off. Theses energies depend on the values of voltage and current being commutated by the power device. As such, the peak motor current and maximum operating voltage of the machine need to be considered. Finally, it can be seen in Eq. 3 that the switching losses are directly proportional to the switching frequency of the power devices. Usually, the switching frequency is a compromise between efficiency and audible noise as lower frequencies tend to generate noise from the machine windings.

$$P = \left(E_{on} + E_{off}\right)\frac{f_s}{2} \tag{3}$$

Besides the characteristics of the power devices (V_{cesat}) and the system switching frequency, the control strategy used for the machine can impact the amount of losses from the inverter stage. Figure 2 depicts the standard space vector pulse width modulation (SVPWM) used to control a 3 phase machine with sinusoidal current. Using this pattern, there are 6 switching events each PWM period. By contrast, the switching pattern in Fig. 3, also known as low loss 2 phase switching PWM, can generate the same current waveforms as the pattern form Fig. 2, but only introduces 4 switching events per PWM period. However, it is important to notice that this technique introduces imbalance in the sharing of losses between upper and lower switches of the inverter stage.

With knowledge of the power devices electrical characteristics, the system switching frequency, and the machine control strategy, it is possible to map the electronic drive losses versus winding current as can be seen in Fig. 4.

Finally, after characterization of the electronics losses, the packaging, mounting options, as well as heat transfer methods are addressed. This information, including thermal grease or thermal pads, needs consideration in thermal simulations.

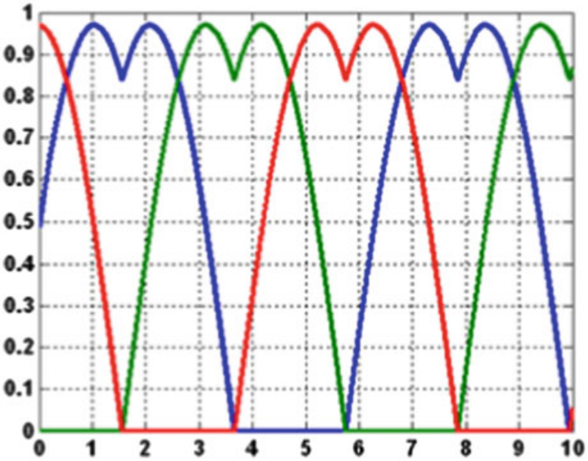


Fig. 3 Low loss PWM(pulse width modulation) switching pattern

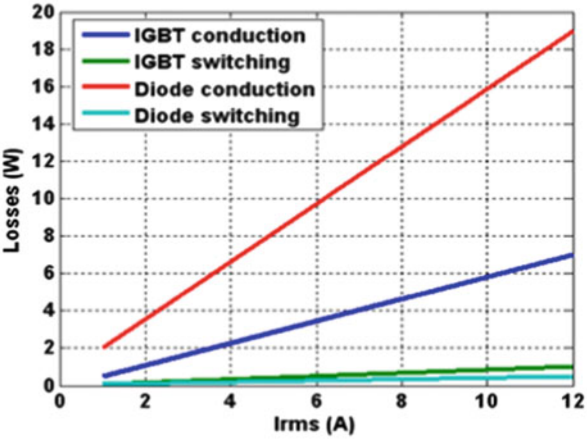


Fig. 4 Inverter losses breakdown

4 CFD Thermal Analysis

After design considerations for thermal losses in the electronics have been accounted for, CFD can be used to evaluate the thermal performance of the finalized electronics layouts and enclosures as well as their system sensitivity. Types of thermal analysis that can be performed through CFD include conduction, convection (forced and natural), phase change, and radiation. Natural convection was considered for the enclosure and conduction for the electronic components.

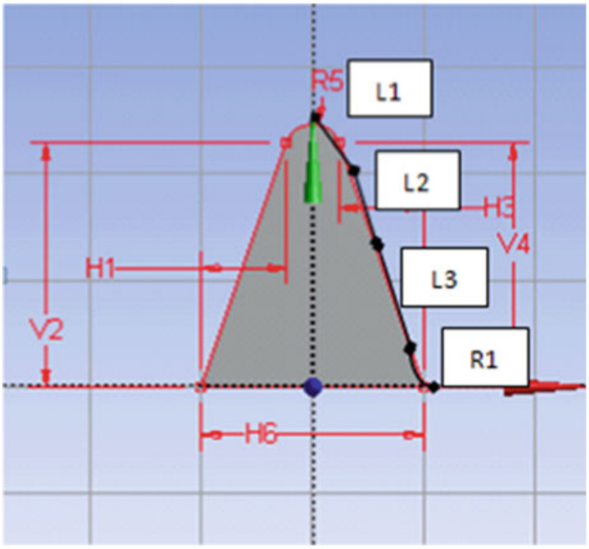


Fig. 5 Cooling fin design parameters

5 Optimization Methods

Optimization methods for thermal management of electronics can focus on proper positioning of the electronics in the enclosure, effective heat sink design, and overall thermal performance of the enclosure within the system. In this study an enclosure design was evaluated in an elevated ambient environment with no forced air over or inside of the enclosure. The individual characteristics of the enclosure were optimized independently and the final enclosure was evaluated to ensure performance requirements were met. More detail will follow regarding the individual fin optimization and the final design of the electronics enclosure.

6 Cooling Fin Optimization Results

A goal driven optimization was performed on a single fin intended to be used on the outer surface of the electronics enclosure. Only natural convection heat transfer was considered. The simulation accounted for full buoyancy and boundary layer effects. The fin material used for the optimization was aluminum, but material variation was also considered to account for material variation and porosity. Steady state analysis was used with the K-Epsilon realizable turbulence model.

The heat generation rate used in the simulation was set for the component with the highest generation rate. The fin geometrical parameters can be viewed below in Fig. 5. The various sections and the bottom radius are the parameters that were optimized. Height was fixed to prevent the trivial solution of an infinitely large fin.

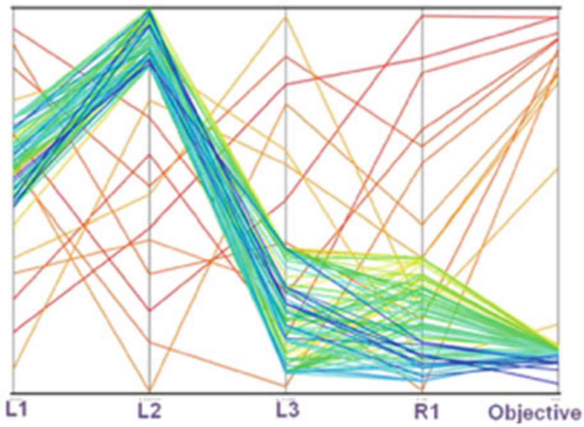


Fig. 6 Fin optimization parameters impact on objective function

The relation of the length in the fin sections to the objective function, optimization of heat transfer, can be viewed in Fig. 6. The optimized geometry would have a second section larger than the first and third sections with a small radius. Three candidate points were compared in the overall electronics enclosure design.

A finalized control cover design using the optimized fin topology was designed. Three dimensional characteristics of the fin design were also analyzed and side cooling channels along the perimeter of the electronics drive enclosure were added to assist natural convection. A validation test was then performed through thermal imaging to confirm CFD results and method.

Many design aspects require consideration for thermal management of electronics cooling including component choice, software, and enclosure thermodynamics. Design and optimization of the individual design characteristics of the electronics enclosure and system have been considered. Future research will focus on full system optimizations using combined physics simulations. Inclusion of radiation models and various material options will also be considered. Sensitivity to ambient conditions and design tolerance will also be addressed.

References

1. Arora, J.S.: Introduction to Optimum Design, 3rd edn. Elsevier (2011).
2. Batarseh, I.: Power Electronic Circuits. Wiley (2004).
3. Davidson, P.A.: Turbulence: An Introduction for Scientists and Engineers (2004).
4. Pletcher, R.H., Tannehillm, J.C., Anderson, D.A.: Computational Fluid Mechanics and Heat Transfer, 3rd edn. CRC Press (2013).
5. White, F.M.: Viscous Fluid Flow, 3rd edn. McGraw-Hill (2011).

A CFD Optimization of Airflow Efficiency for an Electric Motor Driven Centrifugal Fan System for Residential HVAC Applications

Rachele Cocks and Joshua Westhoff

Abstract This study focuses on the design and optimization of an electronics enclosure intended for use in a centrifugal fan being driven by an electric motor for the residential heating, ventilating, and air conditioning(HVAC) market. Typically in these systems the motor is mounted directly in the airstream of the centrifugal fan, but in this case the Regal Beloit's axial motor technology allows for the minimization of this obstruction to the airflow. In the system analyzed the axial motor is mounted in the center of the centrifugal fan and the electronics used to drive the system is enclosed and mounted to the axial motor. This enclosure has been optimized for system airflow efficiency and thermal management of the electronics. A sensitivity analysis was also performed to understand the optimized design's performance under various application environments. Computational fluid dynamics (CFD) was used as a test platform and tool for optimization. The CFD analysis was driven by goal optimization software to explore the design space and lead to an optimized design for overall efficiency. Results were validated to test data and test visualization methods. This presentation will cover the design requirements and details of the application, the optimization and CFD techniques used, and the criteria used for CFD model validation.

1 Introduction

Regal Beloit Corporation has applications and products in the commercial, industrial, and residential air moving markets. These products require a strong focus on efficiency while maintaining current functionality levels. In order to provide increasingly optimized and enhanced product offerings in these markets, Regal Beloit has been using computational tools including computational fluid dynamics (CFD) to support the design process and create innovative products.

R. Cocks (✉) · J. Westhoff
Regal Beloit Corporation, Fort Wayne, IN, USA
e-mail: Rachele.Cocks@RegalBeloit.com

J. Westhoff
e-mail: joshua.westhoff@regalbeloit.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_21

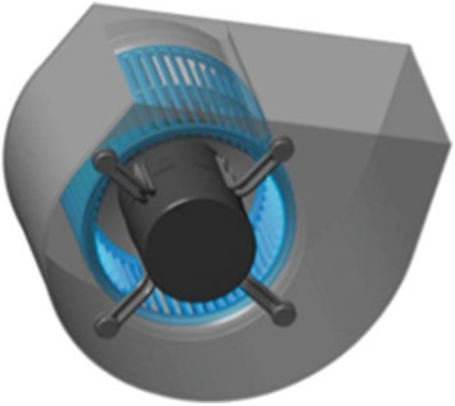


Fig. 1 Blower housing assembly with centrifugal fan, radial motor, and mounting arms and assembly

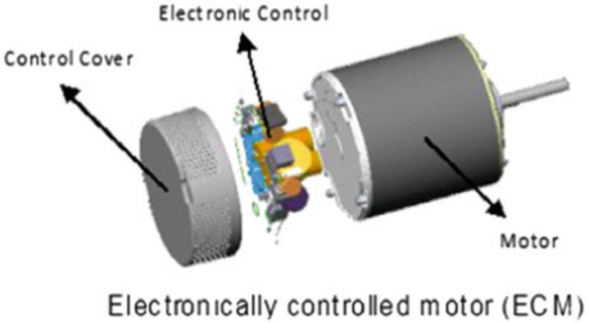


Fig. 2 Radial motor construction with integrated control electronics

2 System Overview and Design Strategy

The system being analyzed and designed in this study is a traditional centrifugal fan blower system used in the residential heating, ventilating, and air conditioning (HVAC) market. An image of a typical blower housing system may be viewed in Fig. 1. The blower is used to force air across a heat exchanger and distribute conditioned air.

The most efficient motors today drive the fan using electronic circuits to control motor operation. These products have been designed similar to the preceding permanent split capacitor technology motors which utilize a radial air gap construction. These motors are designed with a cylindrical form and are assembled with mounting arms to one of the inlet plates of the blower housing with a shaft set-screw to assemble onto the blower wheel. An example of radial motor with integrated control electronics and cover may be viewed below in Fig. 2.

The trade-off for achieving the increased efficiency by using an electronic control is not only the cost but also is the restriction provided by the geometry of its cylindrical form with flat plate cross section. When assembling the motor into one of the inlet orifices, the motor significantly restricts the amount of air entering the fan in the blower

Legislation and overall market demand for improved energy efficiency have been Regal Beloit's drive to continuously improve their air moving products. Keeping with the traditional radial motor form limits the capabilities of the HVAC blower system. There are critical qualities that are required by customers developed as part of the baseline product that must be maintained, the primary of those being serviceability and reliability. Keeping those in mind as well as optimizing the system efficiency, the Regal Beloit engineering team has created a new innovation for the HVAC market

3 Enabling Technology

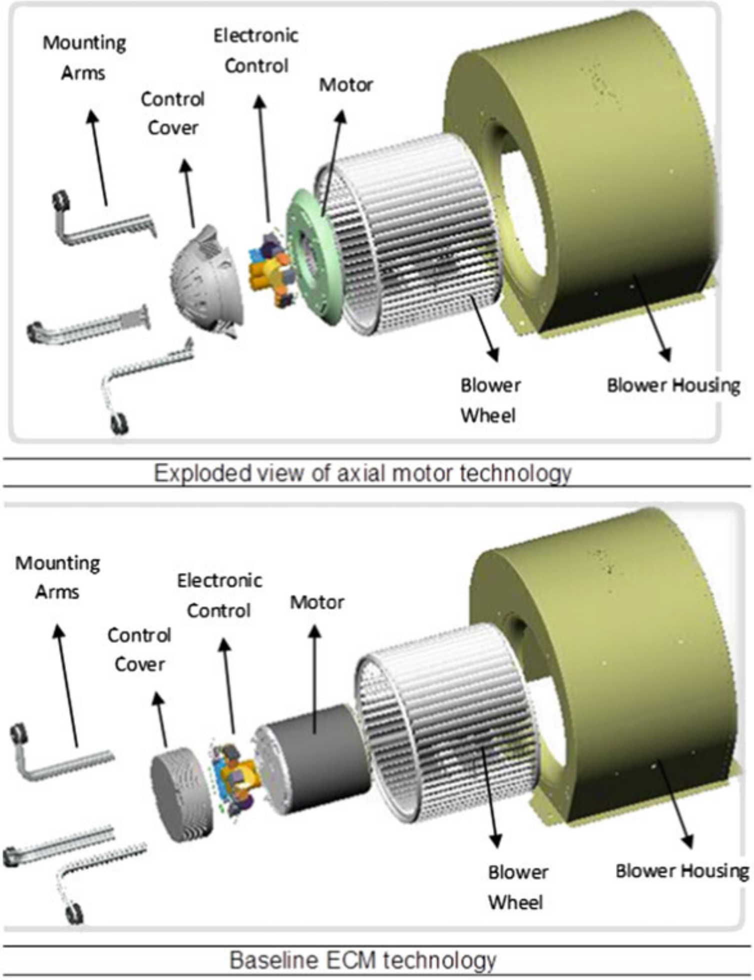
The Regal Beloit engineering team has enabled improvements to blower system efficiency by developing a new design for this application which uses axial flux motor technology. This technology changes the form of the motor to one of a larger diameter, flat plate. Along with Regal Beloit's patented motor construction, the axial flux technology reduces the overall product length to greatly reduce the amount of air flow restriction. A topology comparison of a standard radial motor system to axial motor system may be viewed below in Fig. 3.

Considering the new axial motor form factor, the mechanical construction for the electronic control was redesigned. The layout allows for a new geometric form that provides environmental protection, minimizes air restriction, and improves air flow attachment on its surface to aid in heat transfer and cooling of the motor and electronic components

3.1 CFD Methods and Validation

Various levels of system complexity and fidelity can be modeled to support the innovation and design process. CFD can be used to analyze the thermal and aerodynamic performance of air-moving products such as the previously detailed blower and fan system. The characteristics that enhance thermal performance and aerodynamic performance are not always consistent. This requires various trade off studies to arrive at the right design for the product. In this particular case the thermal analysis was secondary to the aerodynamic performance as the aerodynamic performance is essential to the system efficiency. In order to provide designs that will provide real world performance, complex CFD validation models that simulated an actual test set up were performed with the addition of mesh refinement.

In the CFD validation case, an airflow test based on the AMCA 210 test standard was used and matched to the actual test set up. An example of the simulation model can be viewed in Fig. 4 bellow. Multiple operation points were evaluated, and predicted airflow was within 2 percent.

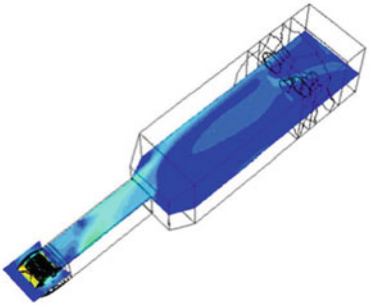


Exploded view of axial motor technology

Baseline ECM technology

Fig. 3 Comparison of radial motor and axial motor construction for centrifugal blower application

Fig. 4 Geometry outline for CFD validation case including airflow test chamber



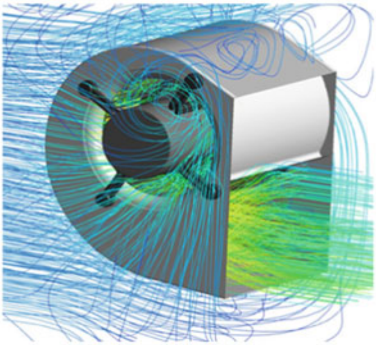


Fig. 5 CFD results for example CFD sub model

The validation case started with a mesh refinement study to evaluate solution dependence on mesh spacing. As the nature of the system is time dependant, a transient simulation was used with a moving reference frame modeling the movement of the centrifugal fan. Prism layers were used on the blades to capture boundary layer effects and the realizable K-Eplson turbulence model was utilized. These results provided satisfactory values and measured trends between configurations were also satisfactory

3.2 Optimization Methods and System Sensitivity.

Sub models using the same mesh spacing and set up as the validation model were created which modeled the blower system using the outlet conditions that were measured at the inlet of the test chamber simulation. This model can be viewed in Fig. 5. This was used to improve the computational efficiency of the simulation model and allow the design team to look at a greater number of tradeoffs in a shorter time. Component level simulations were also used in order to initially optimize the designs before modeling them in the blower system model. Optimization methods included the use of the adjoint method to examine surfaces changed that would minimize force over the control cover and minimize pressure drop through the system. Goal driven optimization of simplified 2D models were also used early on in the design process to evaluate overall control cover topology trends inside of the design space. Special attention was taken to the boundary layer over shape variation.

This study focused on the specific optimization of the electronics control cover positioned in the center of an HVAC centrifugal fan blower system. The control cover restricts and disrupts airflow as it enters radially into the centrifugal fan.

Consideration was also given to the thermal performance of the design in order to ensure adequate cooling of the electronics. Once the optimized design was complete, a sensitivity analysis was performed to evaluate the design's robustness to its application space. Sensitivity to wall proximity, operating points, and ambient conditions were considered. Future research would focus on variation due to manufacturing tolerances.

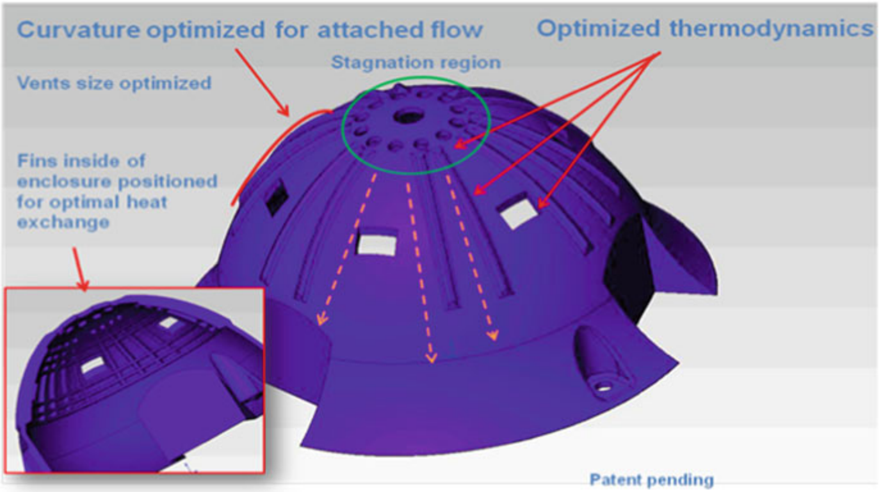

Curvature optimized for attached flow
Optimized thermodynamics
Stagnation region
Vents size optimized
Fins inside of enclosure positioned for optimal heat exchange
Patent pending

Fig. 6 Final optimized electronics control cover

3.3 *Conclusion and Future Research*

The finalized control cover with detailed characteristics can be viewed below in Fig. 6. The openings that provide efficient cooling have been positioned to minimize disruption to the airflow while allowing air into the enclosure to adequately cool the electronics. The overall topology of the cover has been optimized using elliptical forms and the external fins on top of the cover have been designed to enhance cooling while locally directing airflow. The transition at the base of the control cover, which is in close proximity to the entrance of the centrifugal fan, has been designed to prevent flow detachment before entering the fan.

Future research for the optimization of components of the centrifugal fan blower system will focus on optimization at the system level versus component level. Interaction between components will be considered and optimized to further enhance performance. System sensitivity to manufacturing tolerances will also be considered.

References

1. Arora, J. S.: Introduction to Optimum Design, 3rd edn. Elsevier (2011)
2. Batarseh, I.: Power Electronic Circuits. Wiley (2004)
3. Chen, N.: Aerothermodynamics of Turbomachinery Analysis and Design. Wiley (2010)
4. Davidson, P. A.: Turbulence: An Introduction for Scientists and Engineers (2004)
5. Jeffus, L.: Refrigeration and Air Conditioning, 4th edn. Pearson (2004)
6. Pletcher, R. H., Tannehillm, J. C., Anderson, D. A.: Computational Fluid Mechanics and Heat Transfer, 3rd edn. CRC Press (2013)
7. White, F. M.: Viscous Fluid Flow, 3rd edn. McGraw-Hill (2011)

Adoption of New Products with Global and Local Social Influence in a 2D Characteristics Space

M. G. Cojocaru, C. Hogg, C. Kuusela and E. W. Thommes

Abstract We present here an agent-based model (ABM) of adoption of new products including: dynamic consumer preferences and product demands, a 2D characteristics space where products are placed, global and local (nearest neighbours) social influence. The ABM model is built from a continuous time model of the market (Cojocaru et al., Environ Model Softw, 2013), driven by agents' heterogeneity and their local connections. We find that consumer populations where a large fraction of population is sensitive to product popularity displays higher adoption levels of a new product, especially when local social connections are taken into account.

1 Introduction

This chapter looks to better understand the dynamics of social influence on adoption of new product variants ranked by two of their shared characteristics. Consumers are faced with numerous new products today and thus it is of interest to know what influences them to purchase or not, outside of the obvious pricing and quality features. Desirable new products to be adopted are for instance eco-products: variants of known products with environmentally friendly features.

In this work we study the effects of social interaction as a component of new product adoption. Consumers seek advice from friends, and/or consult consumer reviews to help their decision making process. Sociologically this is described as "a process by which an innovation is communicated over time through a social network" [2, 11]. Consumers are often described as: *innovators* and *imitators*. Imitators are influenced by the timing of adoption and decisions made by members of a social network. Innovators feel no social pressure to adopt new products, but rather do it based on the products intrinsic qualities and/or their novelty. Consumer behaviour

C. Hogg and C. Kuusela—at time work was developed.

M. G. Cojocaru (✉) · C. Hogg · C. Kuusela · E. W. Thommes
Mathematics & Statistics, University of Guelph, Guelph, ON, Canada
e-mail: mcojocar@uoguelph.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_22

modelling evolved in several main directions over the past decades, two of which are: continuous time models [2, 8, 9] and discrete models, including individual based ones [3, 6, 7, 15]. Both approaches give valuable information about consumer behaviour on their own. A review of adoption models and their modelling paradigms can be found in [5].

There is important information to be gained from a model that regards the population from a more homogenized point of view, as we do in [5, 12]. Its counterpart, developed as an agent-based model (ABM) here (similarly see [10, 13]), adds new insights based on heterogeneity of consumers and neighbour interactions. We thus model here a dynamic adoption process of a variant ranked via two shared characteristics, with two consumer groups *innovators* and *imitators* using a time-dependent version of the model [1, 3]. We investigate the effects of including local influence and overall popularity of a product variant on the adoption of the variant via sensitivity analysis of the ABM model. The chapter is organized as follows: we present the differentiated product model and our previous continuous time model below. In Sect. 2 we analyze the 2D dynamic adoption model and we introduce local influence over a preferential attachment network among consumers. We close with a few brief remarks.

Differentiated Product Market Model We present the original model as in [1, 11]. A product has m characteristics, giving the characteristics space $\mathbb{R}^m$; n products are placed here at locations $z_1, ..., z_n$ where $z_i = (x_{i1}, ..., x_{im})$, $i = \{1, ..., n\}$ and where $z_j, j > 1$ represent variants of the base product z_1. Consumer preferences are distributed according to a positive density function $f(x)$ where $\int_{\mathbb{R}^m} f(x)dx = N$, the total consumer population. Each consumer purchases the variant that yields them the greatest utility (we assume each consumer buys one of the products). For a consumer located at z, the utility of purchasing variant i is given by:

$$U_i(z) = V_i - \sum_{k=1}^{m} \|z - z_i\|^2, \quad i = 1, ..., n, \tag{1}$$

where $V_i = \alpha_i - p_i$, p_i the price of i, and α_i a quality index of i. V_i can be seen as an objective measure of the value of i, which all consumers agree upon, whereas the second term is the disutility in purchasing a product other then the most preferred one. We define the market space of i as $M_i = \{z \in \mathbb{R}^2 : U_i(z) \geq U_j(z), \quad j \neq i\}$, and define the demand for i as $X_i = \int_{M_i} f(x)dx$

Previous Work In [5] we introduced and analyzed a model of dynamic consumer preferences for two products (base and 1 variant), with social influence, via a partial differential equation model in one characteristic, namely:

$$f_t(z,t) + (v(z,t)f(z,t))_z = \beta f_{zz}(z,t). \tag{2}$$

To ensure the number of consumers remains constant, we imposed boundary conditions: $\int_0^L [f_t + v_z - \beta f_{zz}]dz = 0$ which required $v(0) = v(L)$ and $f_z(0,t) = f_z(L,t)$. We modelled two types of consumers: *Innovators* and *Imitators*, whose preferences evolve differently. Thus, each group has its own preference density so that the total population's density is $f(z,t) = f_{Innov}(z,t) + f_{Imit}(z,t)$. The densities for each group are solutions to a PDE (1.2), with velocities of preference change $v_{Innov} \neq v_{Imit}$, and different coefficients β. The velocities of preference change are taken to be:

$$v_{Innov}(z,t) = z_k \frac{z(z-z_k)(z-L)}{((z-z_k)^2+L)^2};$$
$$v_{Imit}(z,t) = e_1(\sigma)\frac{z(z-z_1)(z-L)}{((z-z_1)^2+L)^2} + e_2(\sigma)\frac{z(z-z_2)(z-L)}{((z-z_2)^2+L)^2}. \tag{3}$$

We consider that v_{Imit} is scaled by a coefficient $e_i(\sigma) = \sigma X_i(z,t) + (1-\sigma)\cdot 1$, $i \in \{1,...,k\}$ where $\sigma \in [0,1]$ is called the social influence parameter. It is the weight an average consumer in *Imit* class places on the popularity of the product, vs. the attributes of the product. The investigation of the 1D PDE model showed the effects of the global social influence, σ, and the proportion of innovators s_{Innov}, on the adoption the variant. The adoption level was higher for either $\sigma \to 0$ and s_{Innov} small (between 2 and 5 %) or $\sigma \to 1$ and s_{Innov} higher (15–20 %).

2 Two Dimensional Adoption Model with Local Social Influence

We can extend our previous work in two ways: a 2D PDE of type (1.2)—currently under investigation— and/or a 2D ABM of the consumer market. Our approach here is to build the 2D ABM model, assuming consumers' preferences evolve independently along each characteristic. This model (implemented in NetLogo) consists of simulating its individuals with a set of parameters representing their attributes: most preferred product, class, intrinsic prefence change, number of neighbours, etc. We keep the two groups as before, *Innov* and *Imit*, we discretize the consumer preference density function over the space of product characteristics and we compute market spaces and demands for the variant. At each time step, a consumer k evaluates its utility of adopting the variant according to (1); the rate of change of its preference is given by solving $\frac{dz_k}{dt} = F_k(z_k,t)$ where $F_{k,j}(z,t) := v_j(z,t) + \beta_k$, $\beta_k := \sqrt{\beta dt}$, $j \in \{Innov, Imit\}$, β drawn from normal distributions with mean 0.05 and standard deviation of 1, and v_j as in (1.3). Initial preference densities are given independently on x and y characteristics: $f_{Imit}(x,0)$, $g_{Imit}(y,0)$ are normal distributions with mean x_1, respectively, y_1 and with variance 0.2 respectively 0.3. Similarly $f_{Innov}(x,0)$, $g_{Innov}(y,0)$ are normal distributions with mean x_2, respectively, y_2 and with standard deviation 1. We next start to investigate this model. We first investigate the effect of the number of consumers on the adoption of the variant. The following parameter values are kept constant throughout the simulations: $L_1 = L_2 := 10$, $(x_1, x_2) = (4,4)$, $(y_1, y_2) = (6,7)$. The second product is introduced on the market at $t_{intro} = 9$ and simulations are run to $T > 0$ (usually $T = 250$) until there is no change in the variant's market share.

Populations where this number is larger display smoother adoption processes as in Fig. 1. The curves have a more concave shape as the social influence coefficients

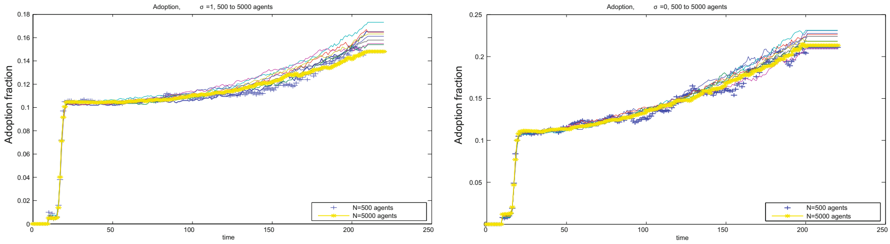


Fig. 1 Adoption with $N \in \{500, ..., 5000\}$, averaged over 80 runs, σ_x, σ_y are 1—*left*, and 0 —*right*

σ_x, σ_y increase (i.e. the imitators weigh the popularity of the products more heavily). Also note that larger σ values lead to higher end-of-time adoption levels in the population. However, as N increases, there are no qualitative differences in the adoption process, thus to investigate the sensitivity to other parameters we run all further simulations with 1500 agents.

Similar to analysis into the 1D model, if here we vary $s_{Innov-x} \in [0.05, 0.2]$ and $\sigma_x \in [0, 1]$, while $\sigma_y = 0.3$ and $s_{Innov-y} = 0.1$ (see Fig. 2—left panel) we see that the regions $\sigma_x \geq 0.6$ and $s_{Innov} \in [0.15, 0.2]$ give higher adoption levels. To further test this scenario, we vary $\sigma_x, \sigma_y \in [0, 1]$ and $s_{Innov-x} \in [0, 0.2]$ (see Fig. 2—right panel) . We see that the region $\sigma_x, \sigma_y \geq 0.5$ and $s_{Innov} \in [0.15, 0.2]$ give higher adoption levels. This again seems to indicate that adoption of the variant is higher in populations where *Innovators* are beyond 10 % (in each characteristic), and the popularity of the products weighs more in the decision of the *Imitators*. It is interesting to see that, unlike the 1D model [5], in the D case, low levels of σ_x, σ_y, and low initial fractions of *Innov* on x lead to low adoption levels.

Local Influence Effects on Adoption of Variants. A further refinement of this model is to consider the consumer population linked over a social network, of a preferential attachment (PA) type [6, 14].

We then consider that a consumer k's immediate link neighbours who adopted the variant exercise an influence over k. We denote one such neighbour by a^{ik}, while a non-adopting neighbour is denoted by n^{ik}. We assume every neighbour's influence is equally weighted by consumer k. To test the effect of the local influence on adoption, we modify the velocity of preference changes of *Imitators* from formula (1.3) to add the terms $\frac{\sum_i a^{ij}}{\sum_i a^{ik}+n^{ij}}\left(\frac{(x_2,y_2)-(x_1,y_1)}{||(x_2,y_2)-(x_1,y_1)||}\right)$.

We set next $s_{Innov-x} = s_{Innov-y} = 0.05$ and vary both global social influence parameters $\sigma_x, \sigma_y \in [0, 1]$. The overall adoption levels with and without the neighbour influence are plotted in Fig. 3 left panel. For varying $s_{Innov-x} = s_{Innov-y} \in [0.05, 0.2]$ and $\sigma_x = \sigma_y = 0.5$, adoption levels with and without local influence are plotted in Fig. 3 right panel.

We see that at low levels of *Innov* on both characteristics, local social influence has a bigger impact (leads to higher adoption levels) whenever the popularity of the products weighs heavily in *Imit* decision ($\sigma_x = \sigma_y$ approach 1). In general,

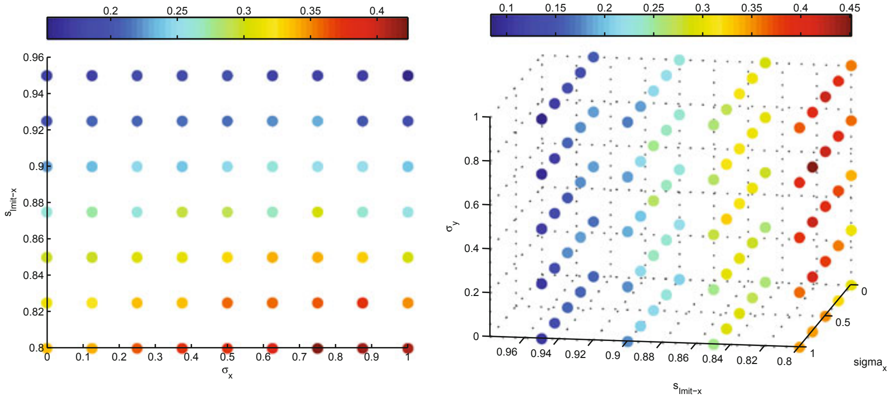


Fig. 2 Adoption levels: lower = *blue*, higher =*red*. *Left* panel shows the adoption for $s_{Innov-x} \in [0.05, 0.2]$ and $\sigma_x \in [0, 1]$, while $\sigma_y = 0.3$ and $s_{Innov-y} = 0.1$. *Right* panel shows the adoption for $s_{Innov-x} \in [0.05, 0.2]$ and $\sigma_x, \sigma_y \in [0, 1]$, while $s_{Innov-y} = 0.1$. The levels were calculated from end-of-time (equilibrium) values over of 400 runs

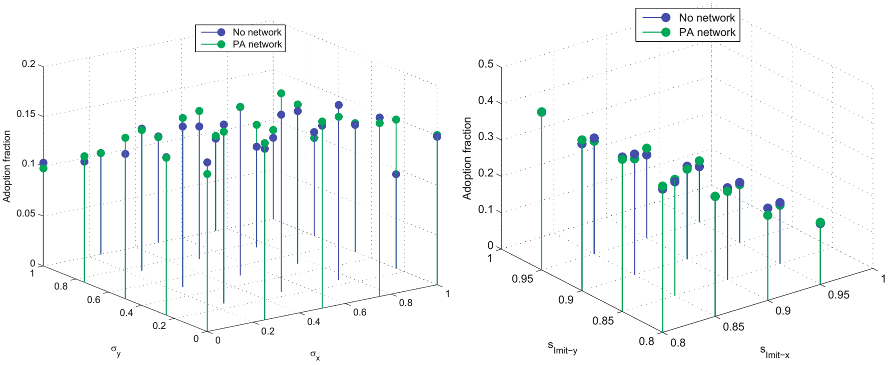


Fig. 3 *Left* panel shows adoption levels for $\sigma_x\sigma_y \in [0, 1]$, $s_{Innov-x}, s_{Innov-y} = 0.05$. *Right* panel represents the adoption fractions for $s_{Innov-x}$, $s_{Innov-y} \in [0.05, 0.2]$ and $\sigma_x = \sigma_y = 0.5$. The levels were calculated from end-of-time (equilibrium) values over 125 runs

for these parameter values, adding the local influence helps the adoption process (albeit by small margins) as seen in Fig. 3. In the case where *Imit* weigh equally popularity of products and their characteristics ($\sigma_x = \sigma_0.5$ Fig. 3 right panel), the local influence does not lead, in general, to a clear increase in adoption levels as we vary the fraction of *Innov* in both characteristics.

Concluding Remarks This chapter shows the first investigations into the behaviour of consumer populations with dynamic preferences in a 2D characteristics space of differentiated products, built from [5]. Its purpose is to identify parameter ranges for global σ and local (neighbour) social influence affecting the adoption of a newer product. We saw that in consumer populations where *Imit* class weighs popularity of products more heavily, local and global social influence work towards increasing the overall adoption levels of the newer product. The sensitivity analysis here is the basis for extended future work on populations with $\sigma_x, y \geq 0.5$, $s_{Innov} \approx 0.05$ and more refined modelling of local social influence weights.

References

1. Anderson, S., de Palma, A., Thisse, J.-F.: Discrete Choice Theory of Product Differentiation. MIT Press (1992)
2. Bass, F. M.: A new product growth model for consumer durables. Manage. Sci. **15**(5):215–227 (1969)
3. Berry, S.T.: Estimating discrete-choice models of product differentiation. R A N D J. Econ. **25**(2):242–262 (1994)
4. Bonabeau, E.: Agent-based modelling: methods and techniques for simulating behaviour. Proc. Natl. Acad. Sci. U S A **99**(Suppl. 3):7280–7287 (2002)
5. Cojocaru, M.-G., Thille, H., Thommes, E., Nelson, D., Greenhalgh, S. Social influence and dynamic demand for new products. Environ. Model. Softw. **50**:169–185 (2013)
6. Delre, S.A., Jager, W., Janssen, M.A.: Diffusion dynamics in small-world networks with heterogeneous consumers. Comput. Math. Organ. Theory **13,** 185–202 (2006)
7. Granovetter, M., Soong, R.: Threshold models of interpersonal effects in consumer demand. J. Econ. Behav. Organ. **7**(1):83–100 (1986)
8. Kalish, S.: A new product adoption model with price, advertising and uncertainly. Manage. Sci. **31**(12):1569–1585 (1985)
9. Mahajan, V., Muller, E., Wind, Y.: New Product Diffusion Models. Kluwer Academic Publishers, Boston (2000)
10. Rahmandad, H., Sterman, J.: Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equations models. Manage. Sci. **54**(5):998–1014 (2008)
11. Rogers, E.M.: Diffusion of Innovasions, 3rd edn. The Free Press (1983)
12. Thille, H., Cojocaru, M.-G., Thommes, E.: A dynamic pricing game in a model of new product adoption with social influence. ASE/IEEE International Conference on Economic Computing (2013)
13. Thommes, E., Thille, H., Cojocaru, M.-G., Nelson, D.: A time-dependent ABM model of an eco-product market with social interactions and dynamic pricing schemes. IEEE Proceedings of TIC-STH, Toronto (2009)
14. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature **393**(6684): 440–442 (1998)
15. Young, R., Peyton, H.: Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. Am. Econ. Rev. **99**:1899–1924 (2009)

On the Group Analysis of a Modified Novikov Equation

Priscila Leal da Silva and Igor Leite Freire

Abstract In this work, we study a modified Novikov equation using group methods. A complete group classification is carried out. Then from the point symmetry generators, we find the one-parameter group of local diffeomorfisms which preserves the equation. From the Lie symmetry generators, we also obtain exact solutions to the considered equation. It is also proved that only one nontrivial conservation law can be established using Ibragimov's recent developments.

1 Introduction

In a recent paper [10] the new integrable equation

$$u_t - u_{txx} + 4u^2 u_x - 3uu_x u_{xx} - u^2 u_{xxx} = 0, \tag{1}$$

with cubic nonlinearities was discovered by V. S. Novikov and is currently known as Novikov equation. Since then, many papers have been dedicated to study different properties of (1). In particular, in [9] it was introduced the *modified* Novikov equation

$$H := u_t - u_{txx} + (b+1)u^2 u_x - buu_x u_{xx} - u^2 u_{xxx} = 0, \tag{2}$$

where b is a real parameter. Clearly such equation generalizes (1).

More recently, Bozhkov, Freire, and Ibragimov studied (1) from the point of view of Lie symmetries. They showed that (1) admits a 5D symmetry Lie algebra. Explicit solutions were obtained. In addition, conservation laws were investigated using recent developments due to Nail Ibragimov in [5, 6, 7]. For further details, see [3].

P. L. da Silva (✉) · I. L. Freire
Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Av. dos Estados, 5001, Santo André, SP, Brazil
e-mail: priscila.silva@ufabc.edu.br

I. L. Freire
e-mail: igor.freire@ufabc.edu.br

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_23

In the present work we carry out a complete group classification of Eq. (2). We show that for any $b \neq 3$, the symmetries are given by the spatial and time translations

$$G_1 : (x, t, u) \mapsto (x + \varepsilon, t, u),\ \ G_2 : (x, t, u) \mapsto (x, t + \varepsilon, u), \tag{3}$$

respectively, and by the dilation

$$G_3 : (x, t, u) \mapsto (x, e^{2\varepsilon} t, e^{-\varepsilon} u). \tag{4}$$

Whenever $b = 3$, in addition to the mentioned symmetries, we have two other transformations preserving solutions:

$$G_4 : (t, x, u) \mapsto \left(t,\ -\frac{1}{2} \ln (e^{-2x} - 2\varepsilon),\ \frac{u}{\sqrt{1 - 2\varepsilon e^{2x}}}\right) \tag{5}$$

and

$$G_5 : (t, x, u) \mapsto \left(t,\ \frac{1}{2} \ln (e^{2x} - 2\varepsilon),\ \frac{u}{\sqrt{1 - 2\varepsilon e^{-2x}}}\right). \tag{6}$$

We observe that the discrete symmetry $(x, t, u) \mapsto (-x, t, u)$ maps the transformation (4) into (6).

Once having these transformations we can easily construct solutions. This is done in Sect. 3. Moreover, in the same section we construct some invariant solutions using the Lie point symmetry generators.

Next, in Sect. 4 we establish conserved currents for the investigated equation. In [3], it was shown that the nonlinear self-adjointness implies in the strict self-adjointness. Then in this chapter, we look for necessary and sufficient condition in order to Eq. (2) be strictly self-adjoint. Then it is obtained a remarkable fact, given by the following

Theorem 1 *Equation (2) is strictly self-adjoint if and only if* $b = 3$.

This theorem will be proved in Sect. 4. Then, in the next, we derive a local conserved current obtained using Ibragimov's approach. The only nontrivial local conservation law established is $C = (C^0, C^1)$, whose components are

$$C^0 = u^2 + u_x^2,\ \ C^1 = 2u^4 - 2u^3 u_{xx} - 2uu_{tx}. \tag{7}$$

Physically speaking, the component C^0 corresponds to the conserved density while C^1 is the conserved flux.

2 Lie Symmetries

Let $x = (x^1, x^2, \ldots, x^n)$ and $u = u(x)$ be, respectively, n independent variable and a smooth function. A Lie point symmetry of the differential equation (DE)

$$F(x, u, u_{(1)}, \ldots, u_{(n)}) = 0, \tag{8}$$

where $u_{(i)}$ denotes the set of derivatives of order i, is a one-parameter transformation group that leaves (8) invariant. For each symmetry one can associate a unique operator

$$X = \xi^i(x,u)\frac{\partial}{\partial x^i} + \eta(x,u)\frac{\partial}{\partial u}, \tag{9}$$

called Lie point symmetry operator. Here, the summation over the repeated indices is presupposed.

A necessary and sufficient condition for (9) to be a Lie point symmetry operator of (8) is

$$X^{(n)}F = \lambda(x,u,u_{(1)},\ldots,)F, \tag{10}$$

where $X^{(n)}$ is the nth prolongation of X, given by

$$X^{(n)} = X + \zeta^{(1)}_{i_1}\frac{\partial}{\partial u_{i_1}} + \ldots + \zeta^{(n)}_{i_1 i_2 \ldots i_n}\frac{\partial}{\partial u_{i_1}u_{i_2}\ldots u_{i_n}},$$

where $\zeta^{(0)} = \eta,\ \zeta^{(k)}_{i_1 i_2 \ldots i_k} = D_i\zeta^{(k-1)} - u_i D_i\xi,\ 1 \le k \le n$, and

$$D_i = \frac{\partial}{\partial x^i} + u_i\frac{\partial}{\partial u} + u_{ij}\frac{\partial}{\partial u_j} + \cdots$$

is the total derivative operator with respect to variable x^i. We guide the diligent reader to [1, 2, 4, 11] for further details.

Consider Eq. (2). Applying condition (10) to (2), and considering the generator

$$X = \xi(x,t,u)\frac{\partial}{\partial x} + \tau(x,t,u)\frac{\partial}{\partial t} + \eta(x,t,u)\frac{\partial}{\partial u},$$

we conclude that $\tau = \tau(t), \eta = \alpha(x,t)u + \beta(x,t), \xi_{xx} = 2\alpha_x$ and

$$2(b+1)u\eta - \xi_t + (b+1)u^2(\eta_u - \xi_x) - bu\eta_{xx} - u^2(3\eta_{xxu} - \xi_{xxx}) +$$
$$-(2\eta_{xtu} - \xi_{xxt}) = \lambda(b+1)u^2,$$

$$\eta_u - \tau_t - \eta_{xxu} = \lambda, \qquad bu\eta_x + 3u^2(\eta_{xu} - \xi_{xx}) + 2\eta_{tu} - 2\xi_{xt} = 0,$$

$$2\eta_{xu} - \xi_{xx} = 0, \qquad 2u\eta + u^2(\eta_u - 3\xi_x) - \xi_t = \lambda u^2,$$

$$\eta + u(2\eta_u - 3\xi_x) = \lambda u,\ \ \lambda = \eta_u - 2\xi_x - \tau_t,\ \ (b+1)u^2\eta_x + \eta_t - u^2\eta_{xxx} - \eta_{txx} = 0$$

The solution of the system reads

$$X_1 = \frac{\partial}{\partial x}, \quad X_2 = \frac{\partial}{\partial t}, \quad X_3 = 2t\frac{\partial}{\partial t} - u\frac{\partial}{\partial u}, \tag{11}$$

if $b \neq 3$, and whenever $b = 3$, along with generators (11),

$$X_4 = e^{2x}\frac{\partial}{\partial x} + e^{2x}u\frac{\partial}{\partial u}, \quad X_5 = -e^{-2x}\frac{\partial}{\partial x} + e^{-2x}u\frac{\partial}{\partial u},$$

reobtaining the 5D Lie algebra found in [3].

Now, in order to obtain the corresponding point transformations, we employ the exponential map. In fact, if X is a Lie point symmetry generator, then

$$e^{\varepsilon X}(x,t,u) := \left(\sum_{j=0}^{\infty}\frac{\varepsilon^j}{j}X^j x, \sum_{j=0}^{\infty}\frac{\varepsilon^j}{j}X^j t, \sum_{j=0}^{\infty}\frac{\varepsilon^j}{j}X^j u\right), \tag{12}$$

where

$$\sum_{j=0}^{\infty}\frac{\varepsilon^j}{j}X^j x = x + Xx + \frac{e^2}{2}X(Xx) + \cdots,$$

Substituting X_1, X_2, X_3, X_4, X_5 into (12), the transformations (1)–(6) are obtained.

3 Invariant Solutions

Here we employ the characteristic method for finding some solutions to (2). For further details, see [1, 2, 4, 11].

For the modified Novikov equation (2), we shall only consider here the construction of solutions for $b \neq 3$ with generators (11). The invariant solutions for $b = 3$ can be found in [3].

Considering $X_3 = 2t\frac{\partial}{\partial t} - u\frac{\partial}{\partial u}$ and using $\xi^1 = 0, \xi^2 = 2t, \eta = -u$ in the characteristic equations

$$\frac{dx}{\xi^1} = \frac{dt}{\xi^2} = \frac{du}{\eta},$$

we obtain the invariant

$$u(x,t) = \frac{\phi(x)}{\sqrt{t}}. \tag{13}$$

Once it has to be a solution of (2), using the respective derivatives, one obtains the ordinary differential equations (ODE) $-\frac{1}{2}\phi + \frac{1}{2}\phi'' + (b+1)\phi^2\phi' - b\phi\phi'\phi'' - \phi^2\phi''' = 0$, which has $\phi(x) = \alpha e^{\pm x}$ as a family of solutions.

Now acting the translational transformation groups generated by X_1 and X_2, we obtain the 3-parameter family of solutions

$$u_{\alpha,\delta,\gamma}(x,t) = \frac{\alpha e^{\pm x + \delta}}{\sqrt{t+\gamma}}.$$

4 Strict Self-Adjointness and Local Conservation Laws

A vector field $C = (C^0, C^1)$ is called conserved vector or conserved current to the Eq. (2) if the following relation holds on the solutions of (2)

$$D_t C^0 + D_x C^1\big|_{F=0} = 0. \tag{14}$$

In [5] Ibragimov states that if Eq. (8) admits a symmetry operator (9), then the quantity

$$C^i = \xi^i \mathcal{L} + W\frac{\delta \mathcal{L}}{\delta u_i} + \sum_{s\geq 1} D_{i_1}...D_{i_s}(W)\frac{\delta \mathcal{L}}{\delta u_{ii_1...i_s}}, \tag{15}$$

with $W = \eta - \xi^i u_i$, $i = 1, \ldots, n$, provides a conservation law for the system

$$\begin{cases} F = 0, \\ F^* = 0, \end{cases} \tag{16}$$

where $\mathcal{L} = vF$ is called *formal Lagrangian*,

$$F^* = \frac{\delta \mathcal{L}}{\delta u}$$

is called *adjoint equation to* (8) and

$$\frac{\delta}{\delta u} = \frac{\partial}{\partial u} + \sum_{j=1}^{\infty}(-1)^j D_{i_1}\cdots D_{i_j}\frac{\partial}{\partial u_{i_1\cdots i_j}}$$

is the Euler–Lagrange operator.

A DE $F = 0$ is said to be nonlinearly self-adjoint if

$$F^*\big|_{v=\phi(x,u)} = \sigma F \tag{17}$$

holds for certain functions ϕ and σ, where this last function may depend on t, x, u and u derivatives. In particular, whenever $\phi = u$, equation $F = 0$ is said to be strictly self-adjoint. For further details, see [5, 6, 7].

Consider the modified Novikov equation (2). Its adjoint equation H^* is given by

$$\begin{aligned} H^* = &-v_t + v_{txx} + v(6-3b)u_x u_{xx} + \\ &+ v_x\left[(-b+6)uu_{xx} - (b+1)u^2 + (6-2b)u_x^2\right] + v_{xx}(-b+6)uu_x + u^2 v_{xxx}. \end{aligned}$$

On the one hand, it is easy to check that when $b = 3$, Eq. (2) is strictly self-adjoint. On the other hand, assuming that the modified Novikov equation is strictly self-adjoint, from the coefficient of u_t we conclude that $\sigma = -1$ and then $b = 3$. This not only proves Theorem 1 but also implies that Ibragimov's theorem can only be used to find local conservation laws for (2) if $b = 3$.

In [3] it was proved that the nonlinear self-adjointness of the Eq. (1) is equivalent to the strict self-adjointness. For this reason it is sufficient to study the strict self-adjointness of Eq. (1).

Consider the generator X_3 and $\mathcal{L} = vH$. Then using $W = -u - 2tu_t$ and the respective derivatives of $\mathcal{L}$ in (15), we obtain

$$\tilde{C^0} = u^2 + u_x^2 + D_x\,(B)\,,\;\; \tilde{C^1} = 2u^4 - 2u^3u_{xx} - 2uu_{tx} - D_t\,(B),$$

where $B = \frac{4}{3}tuu_{tx} - \frac{2}{3}tu_tu_x - \frac{2}{3}uu_x + 2tu^3u_{xx} - 2tu^4$.

We can eliminate the terms $D_t(-B)$ and $D_x(B)$ since it correspond to a null divergence. Therefore, the components of the conservation law C are given by (7).

By a similar calculation, one can prove that the conservation laws for generators X_1, X_2, X_4, and X_5 are trivial, i.e., $C = (0, 0)$.

Acknowledgement The authors would like to thank FAPESP for financial support, project numbers 2012/22725-4 and 2011/19089-6.

References

1. Bluman, G.W., Anco, S.: Symmetry and Integration Methods for Differential Equations. Springer, New York (2002)
2. Bluman, G.W., Kumei, S.: Symmetries and Differential Equations (Applied Mathematical Sciences 81). Springer, New York (1989)
3. Bozhkov, Y., Freire, I.L., Ibragimov, N.H.: Group analysis of the Novikov equation. Comp. Appl. Math. **33**, 193–202 (2014)
4. Ibragimov, N.H.: Transformation Groups Applied to Mathematical Physics. D. Reidel Publishing Co., Dordrecht (1985) (Translated from the Russian Mathematics and its Applications (Soviet Series))
5. Ibragimov, N.H.: A new conservation theorem. J. Math. Anal. Appl. **333**, 311–328 (2007)
6. Ibragimov, N.H.: Nonlinear self-adjointness and conservation laws. J. Phys. A: Math. Theor. **44**, 432002, 8 pp. (2011)
7. Ibragimov, N.H.: Nonlinear self-adjointness in constructing conservation laws. Arch. ALGA, **7/8,** 1–90 (2011) (See also arXiv:1109.1728v1[math-ph], pp. 1–104) (2011)
8. Ibragimov, N.H., Khamitova, R.S., Valenti, A.: Self-adjointness of a generalized Camassa-Holm equation. Appl. Math. Comp. **218**, 2579–2583 (2011)
9. Mi, Y., Mu, C.: On the Cauchy problem for the modified Novikov equation with peakon solutions. J. Diff. Equ. **254**, 961–982 (2013)
10. Novikov, V.S.: Generalizations of the Camassa-Holm equation. J. Phys. A: Math. Theor. **42**, 342002, 14 pp. (2009)
11. Olver, P.J.: Applications of Lie Groups to Differential Equations. Springer, New York (1986)

Implication of Stochastic Resonance on Neurological Disease Quantification

T. K. Das, N. Rajakumar and M. Jog

Abstract This presents an application of stochastic resonance in a data-driven nonlinear bistable system, in which inhibitory and excitatory electrophysiological neuronal activity in the prefrontal cortex (PFC) is quantified in a control and a putative rodent model of schizophrenia brains. An empirical mode decomposition protocol was applied for processing and analyzing the spike data. Within the different experimental conditions, we extracted different asymmetric shapes of bistable model potentials using the Fokker–Planck equation (FPE). Our analyses in control brains suggest that neuronal firing, along with noise (e.g., synaptic activity) before and after amphetamine administration provide asymmetries with phase transition in the bistable model allowing bidirectional information flow. Such transitions appear to be impaired in the disease model.

1 Introduction and heoretical oncepts

Naturally governed stochastic resonance (SR) exists on the basis of cooperative behavior between "noise" and "nonlinear dynamics" and acquires an enhanced sensitivity in the presence of any small internal/external time dependent forcing [1, 2]. Despite decades of research on "SR" and its widespread multidisciplinary applications, including biological systems (e.g., gene expression, neural systems, etc.) [36], the positive role of "noise" in any neuronal network (either healthy or disease case) for evolving the brain mechanism invivo during "neural encoding/decoding" is mostly ignored by neuroscientists [7]. The mechanism of "SR" is important in this scheme in [3] that it describes the possible occurrence of asymmetry inside the

T. K. Das (✉) · M. Jog
Department of Clinical Neurological Sciences, London Health Sciences Centre,
Western University, N6A 5A5, London, ON, Canada
e-mail: tdas2@uwo.ca

N. Rajakumar
Department of Anatomy and Cell Biology, Western University, N6A 5C1, London, ON, Canada,
e-mail: nrajakum@uwo.ca

M. Jog
e-mail: Mandar.Jog@lhsc.on.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_24

double-well potential when any time-dependent forcing exists and probable information flow can occur between two different dynamic states with noise. Since noise may in fact be an important contributor to understanding neuronal firing, the “SR” phenomena needs to be studied and such analysis methods need to be applied to neuronal data.

The positive (hallucinations, thought disorders) and negative symptoms (apathy, withdrawal) and cognitive deficits in schizophrenia have a significant impact on the patient. However, mechanisms underlying negative and cognitive symptoms are not clear. An important contribution to this problem may include the lack of adequate animal models to investigate these negative symptoms. Studies in human and non-human primates have suggested that a possible origin of the cognitive and negative symptoms may be within the prefrontal cortex (PFC) [8]. A rodent model has been developed using human recombinant nerve growth factor (hrNGF) injections in the neonates, showing adult onset dopaminergic hyperactivity, social interaction deficits and a number of structural features described in postmortem brains of patients with schizophrenia [9, 10]. A histological study on the hrNGF rat model shows that partial ablation of subplate and GABAergic synaptic abnormalities of the PFC are responsible in altering dopaminergic activity (DA) [11]. However, electrophysiological properties of this model have not been adequately studied.

Recently, using “SR,” Zheng et. al. presented a bistable model in order to explore bimodality in stochastic gene expression with additive and/or multiplicative external noise [6]. In contrast to earlier studies, we applied a significantly modified version of the Zheng methodology to extract bimodal distribution functions from invivo electrophysiological data to study the stochastic bistable potential wells of the Fokker–Planck equation (FPE) [12] in the control and hrNGF rodent model.

2 Experimental Methodology

A group of seven Sprague Dawley male rodent pups received neonatal injections of hrNGF into the developing PFC on postnatal day 1. Another group of seven male pups received identical injections of saline in parallel (control). All experiments (animals 600 g each) were carried out under approval of UWO animal ethics committee. Through the EthoVision behavioral monitoring system at 14 weeks of age, the hrNGF lesioned rats demonstrated significantly reduced social interaction, compared to the control group rats [11]. These animals underwent in vivo electrophysiological recordings using accepted methods [13, 14]. Twelve tetrodes for extracellular recordings were independently inserted and moved until the PFC (AP 2.2, L 0.8, DV 3.4 mm from Bregma) is reached. Followed by the baseline recording at10 min, an excitatory type stimulation, d-amphetamine (AMPH) was injected intraperitoneally in both control and hrNGF rodents. Electrophysiological recordings were accomplished for 3 min duration at the time moments 15, 60, and 180 min after injection.

3 Numerical Methodology

3.1 Data Analysis

Current state-of-the-art of data analysis uses either timeseries based analysis or frequency based Fourier transformation and/or timefrequency analysis based on the Hilbert transform or the complex wavelet transform of the signals [15]. All these methods have shown limited applications in the linear and/or stationary regime due to the lack of mathematical rigor. Here, we applied an adaptive multivariate Hilbert–Huang ransform (HHT) that mostly preserves transient and nonlinear features of the data. Our overall data analysis can be divided into two steps(1) the empirical mode decomposition (EMD) and (2) the Hilbert transformation. Followed by an established method [15, 16], the process of EMD starts with proper data sampling on m-hyperspheres using Quasi-Monte Carlo (QMC) based low-discrepancy sequences. This iterative process allows the decomposition of spike signal locally and differentiates the input signal into a finite set of zero-mean "Intrinsic-Mode-Functions (IMF)" components. It is important to mention that the frequency of oscillations decreases with increasing the number of IMFs until it reaches to a residue signal. Each IMF fulfills the requirement of analytic quadrature of any input signal, and a Hilbert ransform of individual IMF components is eventually applied to extract instantaneous quantities like amplitudes (A), phases (Φ) and frequencies (Ω).

3.2 Mathematical Model Analysis

To investigate the characteristics of stochastic resonance on spatiotemporal aspects of neuronal dynamics in presence/absence of external stimulations, any experimental observable "Φ" with possible inherent noise contribution satisfies the Langevin equation

$$\frac{\partial \Phi}{\partial t} = -\frac{\delta \mathrm{U}(\Phi)}{\delta \Phi} + \Phi \gamma(t) + \epsilon(t) \tag{1}$$

Here U(Φ) denotes the stochastic potential function, describing the states of neuronal dynamics in PFC. Also, γ(t)and ϵ(t) define the contribu of Gaussian white noise in multiplicative and additive ways, which has zero mean and correlations

$$\begin{aligned} <\gamma(t)\gamma(t') >&= 2\alpha_m \delta(t-t') \\ <\epsilon(t)\epsilon(t') >&= 2\alpha_a \delta(t-t') \\ <\gamma(t)\epsilon(t')> =<\epsilon(t)\gamma(t') >&= 2\beta\delta(t-t') \end{aligned} \tag{2}$$

where α_m and α_a represent multiplicative and additive noise intensities, respectively. Also β measures the strength of correlation between these two types of noise.

Due to the random fluctuation in the experimental phase variable ($\varPhi$), the probability distribution function, W($\varPhi, t$) of such variable is calculated, which satisfies the associated FPE [12] of Eq. (1) as

$$\frac{\partial W(\varPhi,t)}{\partial t} = -\frac{\partial}{\partial \varPhi}\{A_1(\varPhi)W(\varPhi,t)\} + \frac{\partial^2}{\partial \varPhi^2}\{A_2(\varPhi)W(\varPhi,t)\} \tag{3}$$

with corresponding coefficients as

$$\begin{aligned} A_1(\varPhi) &= -\frac{\delta \mathrm{U}(\varPhi)}{\delta \varPhi} + \alpha_m \varPhi + 2\beta\sqrt{\alpha_m \alpha_a} \\ A_2(\varPhi) &= \alpha_m \varPhi^2 + 2\beta\sqrt{\alpha_m \alpha_a}\varPhi + \alpha_a \end{aligned} \tag{4}$$

In the steady state, the solution of FPE governs the form

$$W_s(\varPhi) = B\exp\{-\mathrm{U}(\varPhi)\} \tag{5}$$

where B is an arbitrary constant and is normalized to be unity. In contrast to earlier studies on stochasticity in gene expression [6], the corresponding potential function, U($\varPhi$) that includes all possible noise contributions has been extracted from our experimental data driven steady state probability distribution function, W$_s$($\varPhi$).

4 Results and Discussion

In the control case, depicted bimodal distributions in baseline, recorded from PFC, are mostly localized in the range of phase, $0 \le \varPhi \le 0.03$, as in Fig. 1a. The resultant asymmetric double-well potential function, which is calculated from Fig. 1a, is shown in Fig. 2a for characterizing stochastic resonance via phase transitions (or switching). The mode location at low $\varPhi$-value shows stronger distribution and more stable synchronized dynamic state than that of its distribution and dynamic state at high $\varPhi$-value. This could be interpreted as weak connectivity in the PFC network, which could open in order to exchange information between these bistable states. After 15 min of giving excitatory type perturbation (AMPH), a strong phase-shift with the switching mechanism of synchronized states is observed (see Figs. 1b and 2b). This symmetry breaking transition could be occurring due to excitatory type perturbation resulting in self-organization of spontaneous neuronal activities. An important component adding to this is the increased amount of noise produced by nonlinear interactions of large number of dopaminergic synapses invivo. The bimodal probability distribution as well as its respective potential function remain the same at the 60 min recording with previous existing phase localiations (see in Figs. 1c and 2c). However, shapes of W($\varPhi$) and U($\varPhi$) at 180 min recording are intended to relax towards the shape of baseline state (near equilibrium) with shifting phases (see Figs. 1d and 2d). This could happen due to dominant cooperation between balanced excitatory and inhibitory neuronal activities in PFC rather than just competition

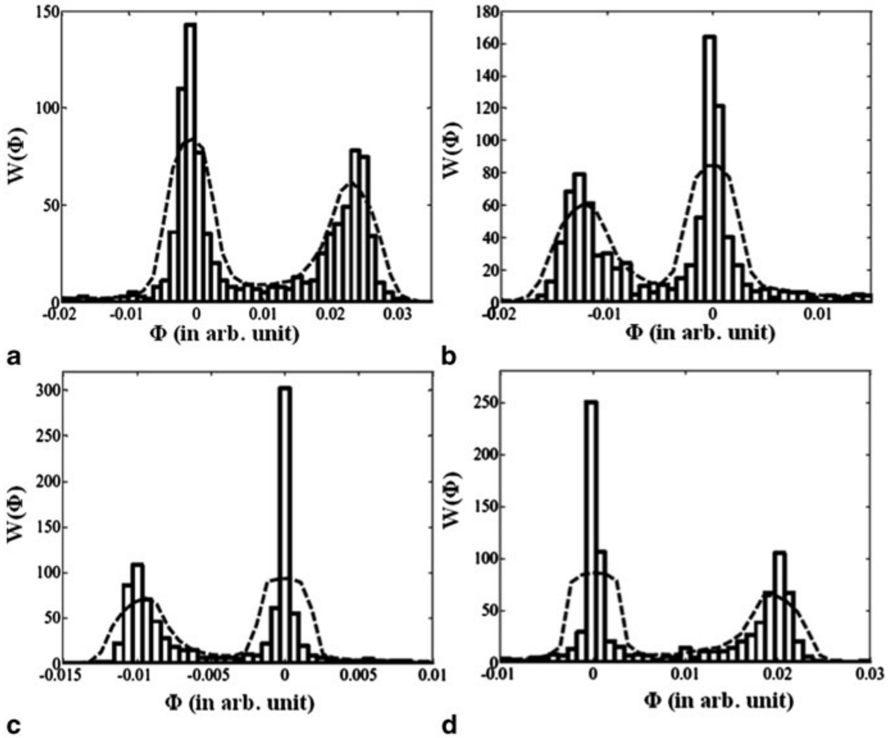


Fig. 1 Phase probability distribution, W(Φ) versus phase, Φ from a multichannel tetrode spike recordings in prefrontal cortex (PFC) of control rats at recording labels **a** baseline **b** 15 min **c** 60 min and **d** 180 min after injecting amphetamine (AMPH)

itself. Besides, the metastable dynamic phase at 15 and 60 min recordings, which is induced by excitatory type perturbation (AMPH), could be conjectured to represent the temporary coordination based memory formation in the healthy control brain.

On the other hand, the bimodal distribution function W(Φ) and the potential function U(Φ), calculated in the baseline recording of hrNGF groups, are populated in the range of phase, $-0.3 \leq \Phi \leq 0$ (see Figs. 3a and 4a). Moreover, the location of mode distribution at high Φ-value is found to be stronger, compared to its distribution at low Φ-value. The corresponding highly stable synchronized dynamic state at higher phase value and weakly synchronized dynamic state at low phase could be interpreted as the rigid (strong) network connectivity in the hrNGF case due to increased GABAergic synaptic abnormalities in the thalamocortical pathways. The resultant phase distribution functions and the potential functions at 15, 60, and 180 min are shown in Fig. 3b, c, d and in Fig. 4b, c, d respectively. Inability of symmetry breaking bistable dynamic states in presence of excitatory perturbation could be due to excessive GABAergic synaptic abnormalities and loss of dopamine fiber densities in

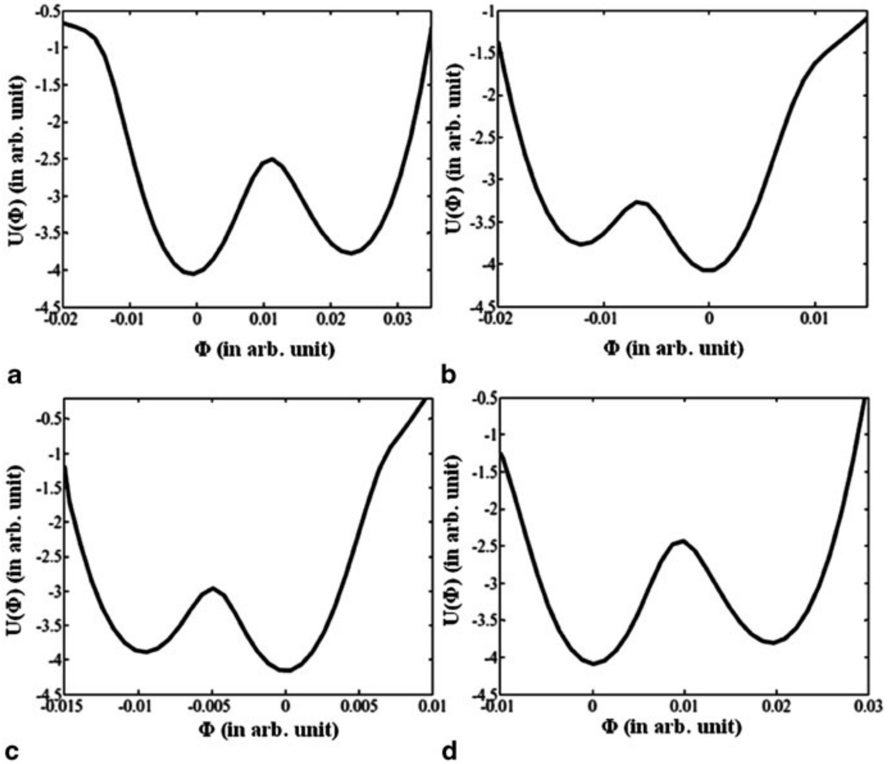


Fig. 2 Calculations of corresponding bistable potential function, U(*Φ*) as a function of phase, *Φ* from a multichannel tetrode spike recordings in prefrontal cortex (PFC) of control rats at **a** baseline **b** 15 min **c** 60 min and **d** 180 min after injecting AMPH

hrNGF model [11]. This could imply unidirectional information flow with/without excitatory perturbation.

5 Conclusions

Our analyses on spatiotemporal aspects of neuronal dynamics suggest that spontaneous neuronal activity, resulting from internally driven neuronal network force (such as synaptic connectivity) and source of physical noise, may provide asymmetries in our bistable model potential that has a strong influence on stochastic resonance effect in order to quantify bistable brain dynamic phase transitions. These may help to characterize dopamine agonist and/or antagonist activities and to demonstrate inability of bidirectional information flow in hrNGF model. The reduced stochastic resonance in dynamic phase space and its relation to potential GABAergic synaptic overactivity in disease states may represent hypofrontality [17] in the hrNGF rat model. Our

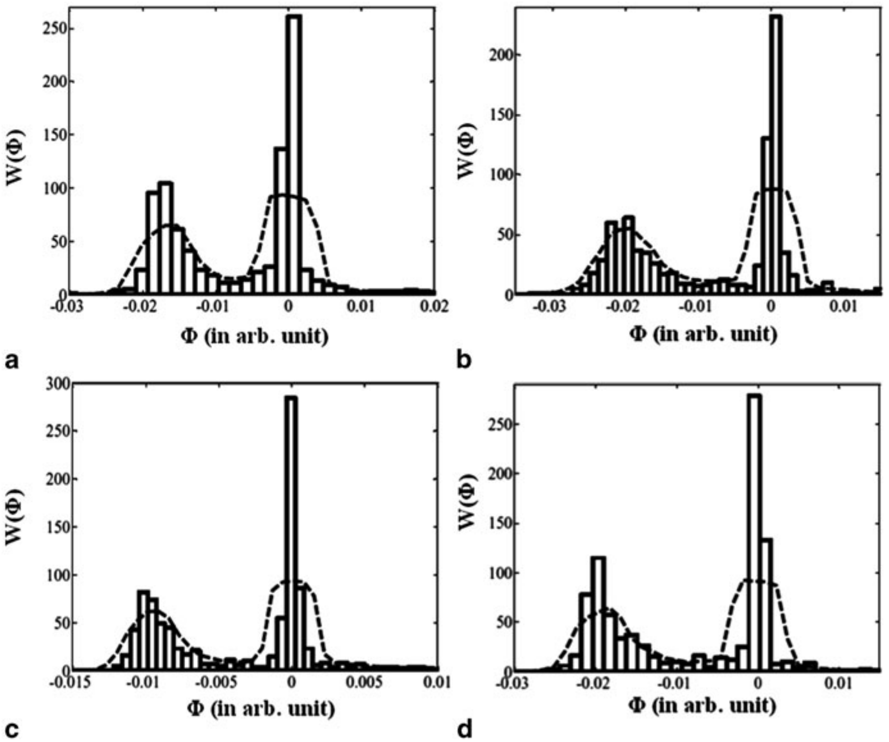


Fig. 3 Same as in Fig. 1, but from a multichannel tetrode spike recordings in prefrontal cortex (PFC) of human recombinant nerve growth factor (hrNGF) rats

analyses in Sect.4 showed that the use of AMPH alone, which increased dopamine levels, was not enough to improve the dynamic brain state of this hypofrontal rat model, and evidence against the dopaminergic hyperactivity as the sole contributor. Experiments with coadministration of GABAergic inhibitors and AMPH may help to mimic the corrected dynamic state as well as the characterization of "negative symptoms" in our hrNGF rodent model. Also, our analyses in control and disease model suggested that the degree of spontaneous symmetry breaking could be quantified as severity of neurological disease. Further study on dynamic multistabilities and corresponding phase transitions may shed light on decision making in the PFC, which may be absent in the disease states.

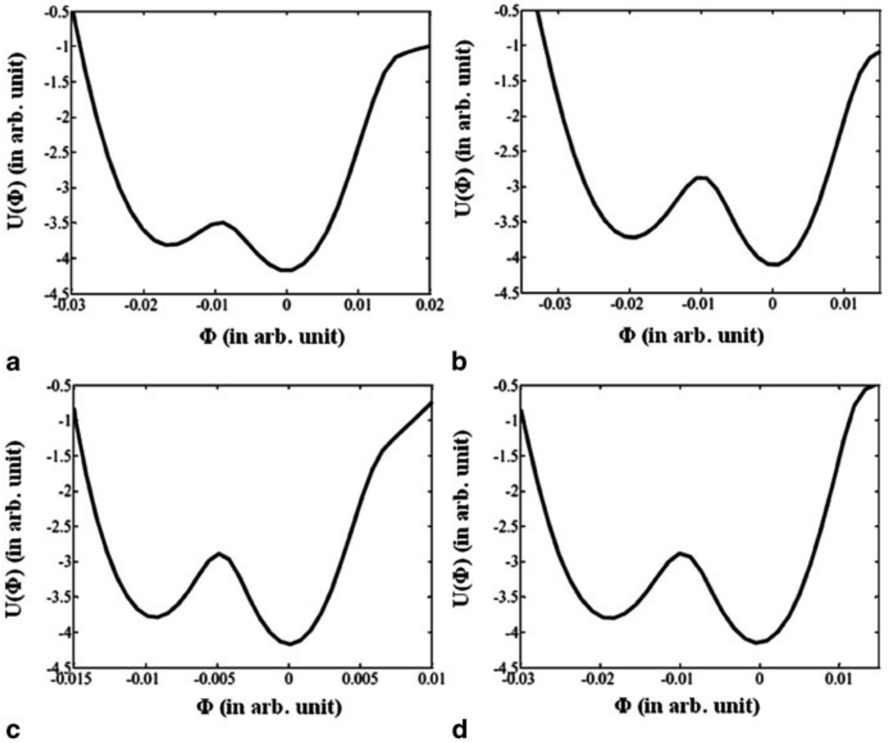


Fig. 4 Same as in Fig. 2, but from a multichannel tetrode spike recordings in refrontal cortex (PFC) of human recombinant nerve growth factor (hrNGF) rats

References

1. McNamara, B., Wiesenfeld, K.: Theory of stochastic resonance. Phys. Rev. A **39**(9), 4854 (1989)
2. Wiesenfeld, K., Moss, F.: Stochastic resonance and the benefits of noise: From ice ages to crayfish and SQUIDs. Nature **373**(6509), 33–36 (1995)
3. Gammaitoni, L., Hänggi, P., Jung, P., Marchesoni, F.: Stochastic resonance. Rev. Mod. Phys. **70**(1), 223 (1998)
4. Swain, P. S., Longtin, A.: Noise in genetic and neural networks. Chaos: Interdiscip. J. Nonlinear Sci. **16**(2), 026101 (2006)
5. Wang, Z., Hou, Z., Xin, H.: Internal noise stochastic resonance of synthetic gene network. Chem. Phys. Lett. **401**(1), 307–311 (2005)
6. Zheng, X.D., Yang, X.Q., Tao, Y.: Bistability, probability transition rate and first-passage time in an autoactivating positive-feedback loop. PloS ONE **6**(3), e17104 (2011)
7. McDonnell, M.D., Abbott, D.: What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. PLoS Comput. Biol. **5**(5), e1000348 (2009)
8. Goldman-Rakic, P.S.: Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Prog. Brain Res. **85**, 325–336 (1990)
9. Bunney, W. E., Bunney, B. G.: Evidence for a compromised dorsolateral prefrontal cortical parallel circuit in schizophrenia. Brain Res. Rev. **31**(2), 138–146 (2000)

10. Cowan, W. M., Harter, D. H., Kandel, E. R.: The emergence of modern neuroscience: Some implications for neurology and psychiatry. Ann Rev. Neurosci. **23**(1), 343–391 (2000)
11. Lazar, N. L., Rajakumar, N., Cain, D. P.: Injections of NGF into neonatal frontal cortex decrease social interaction as adults: A rat model of schizophrenia. Schizophr. Bull. **34**(1), 127–136 (2008)
12. Gardiner, C.W.: Handbook of Stochastic Methods, 3rd edn., pp. 342–372. Springer-Verlag, Berlin (2004)
13. Aur, D., Jog, M. S.: Building spike representation in tetrodes. J. Neurosci. Methods **157**(2), 364–373 (2006)
14. Jog, M. S., Connolly, C. I., Kubota, Y., Iyengar, D. R., Garrido, L., Harlan, R., Graybiel, A. M.: Tetrode technology: Advances in implantable hardware, neuroimaging, and data analysis techniques. J. Neurosci. Methods **117**(2), 141–152 (2002)
15. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. A: Math. Phys. Eng. Sci. **454**(1971), 903–995 (1998)
16. Rehman, N., Mandic, D.P.: Multivariate empirical mode decomposition. Proc. R. Soc. A: Math. Phys. Eng. Sci. **466**, 1291–1302 (2010)
17. Marek, G.J., Behl, B., Bespalov, A.Y., Gross, G., Lee, Y., Schoemaker, H.: Glutamatergic (N-methyl-D-aspartate receptor) hypofrontality in schizophrenia: Too little juice or a miswired brain? Mol. Pharmacol. **77**(3), 317–326 (2010)

Impact of Excess Mortality on the Dynamics of Diseases Spread by Ectoparasites

Attila Dénes and Gergely Röst

Abstract In this chapter, we generalize our earlier model for the spread of ectoparasites and diseases transmitted by them by including disease-induced mortality. The qualitative behavior of the system is similar to that of the original model: three reproduction numbers determine which of the four possible equilibria is globally asymptotically stable. We conclude that a moderate mortality decreases the size of the population, while a high mortality leads to the eradication of the infection. The main tools used for the proofs include persistence theory, Lyapunov–LaSalle theory and Dulac's criteria.

1 Introduction, Basic Properties of the Model

Ectoparasites are present in several regions of the world. Besides the problems caused by the infestation, they are also responsible for the transmission of several diseases like relapsing fever or murine typhus (for details see, e.g., [1]). The spread of these diseases is different from other vector-borne diseases, as in this case, the vectors themselves are transmitted like a disease through the human contact network. In [2], we established a basic model for the spread of ectoparasites and diseases transmitted by them and completely described the global dynamics of the model. Our basic model does not include disease mortality, however, as several ectoparasite-borne diseases are lethal (e.g., epidemic typhus or plague), it is a natural question to ask what happens if we also incorporate disease-induced mortality. In this chapter, we study the model with disease-induced mortality showing that the modified system has a similar behavior as the original one. Some of the proofs in [2] can be applied in an analogous way, however, several of them need some additional ideas or completely different methods.

A. Dénes (✉) · G. Röst
Bolyai Institute, University of Szeged, Aradi vértanúk tere 1., 6720 Szeged, Hungary
e-mail: denesa@math.u-szeged.hu

G. Röst
e-mail: rost@math.u-szeged.hu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_25

The present model is for one ectoparasite species, which might be a vector for a disease as well. The population is divided into three compartments: susceptibles (i.e., those who are not infested, denoted by $S(t)$), those who are infested by noninfectious parasites ($T(t)$) and those who are infested by infectious parasites ($Q(t)$). In the following, we will call an individual from compartment S (resp. T, Q) an S- (resp. T-, Q-) individual. A T-individual might infest an S-individual with noninfectious parasites, while a Q-individual might infest an S-individual or a T-individual with parasites which carry the disease. We assume that a person is infected by the disease, if and only if, he is infested by infectious parasites. We denote the transmission rate from Q to S and T by β_Q, while β_T stands for the transmission rate from T to S. The disinfestation rate is denoted by θ for compartment T and by μ for compartment Q. We denote by b the recruitment and removal rate, and d denotes disease-induced mortality. With these assumptions we obtain the following system of differential equations:

$$\begin{aligned}
S'(t) &= -\beta_T S(t)T(t) - \beta_Q S(t)Q(t) + \theta T(t) + \mu Q(t) + b - bS(t),\\
T'(t) &= \beta_T S(t)T(t) - \beta_Q T(t)Q(t) - \theta T(t) - bT(t),\\
Q'(t) &= \beta_Q S(t)Q(t) + \beta_Q T(t)Q(t) - \mu Q(t) - bQ(t) - dQ(t).
\end{aligned} \tag{1}$$

It is easy to see that all solutions are bounded and solutions with nonnegative initial values remain nonnegative.

Letting $S^* = \frac{(b+d)\theta - b\mu + b\beta_Q}{(b+d)\beta_T}$, the four equilibria can be calculated as:

$$\begin{aligned}
E_S &= (1,0,0), & E_{QT} &= \left(S^*, \tfrac{b+d+\mu}{\beta_Q} - S^*, \tfrac{b\left(\beta_Q-(b+d+\mu)\right)}{(b+d)\beta_Q}\right),\\
E_T &= \left(\tfrac{b+\theta}{\beta_T}, 1-\tfrac{b+\theta}{\beta_T}, 0\right), & E_Q &= \left(\tfrac{b+d+\mu}{\beta_Q}, 0, \tfrac{b\left(\beta_Q-(b+d+\mu)\right)}{(b+d)\beta_Q}\right).
\end{aligned}$$

By introducing a single infested, respectively, infested and infected individual into one of the equilibria E_S, E_T, and E_Q, we obtain three different reproduction numbers. By introducing a T-, resp. Q-individual into E_S, we get the reproduction numbers

$$R_1 = \frac{\beta_T}{b+\theta}, \text{ resp. } R_2 = \frac{\beta_Q}{b+d+\mu}. \tag{2}$$

If we introduce a Q-individual into E_T, we get the same reproduction number R_2 again. Finally, by introducing a T-individual into E_Q, we obtain the reproduction number

$$R_3 = \frac{\beta_T(b+d)(b+d+\mu)}{\beta_Q(b(\beta_Q+\theta-\mu)+d\theta)}. \tag{3}$$

The following proposition can easily be checked.

Proposition 1 *Equilibrium E_S always exists. Equilibrium E_T exists if and only if $R_1 > 1$. Equilibrium E_Q exists if and only if $R_2 > 1$. Equilibrium E_{QT} exists if and only if $R_2 > 1$ and $R_3 > 1$.*

Proposition 2 *Local stability of the four possible equilibria is determined by the reproduction numbers in the following way.*

(i) E_S *is locally asymptotically stable (LAS) if* $R_1 < 1$ *and* $R_2 < 1$*, and unstable if* $R_1 > 1$ *or* $R_2 > 1$.
(ii) E_T *is LAS if* $R_1 > 1$ *and* $R_2 < 1$*, and unstable if* $R_2 > 1$.
(iii) E_Q *is LAS if* $R_2 > 1$ *and* $R_3 < 1$*, and unstable if* $R_3 > 1$.
(iv) E_{QT} *is LAS if* $R_2 > 1$ *and* $R_3 > 1$ *(i.e., always when it exists).*

Proof *(i)* Calculating the eigenvalues of the Jacobian of the linearized equation around the equilibrium E_S we obtain $\lambda_{S_1} = -b, \lambda_{S_2} = -b-\theta+\beta_T = (b+\theta)(R_1-1)$, and $\lambda_{S_3} = -b-d-\mu+\beta_Q = (b+d+\mu)(R_2-1)$. All of the eigenvalues are negative if $R_1 < 1$ and $R_2 < 1$, while at least one of them is positive if $R_1 > 1$ or $R_2 > 1$.

(ii) If we linearize around the equilibrium E_T, we find the eigenvalues $\lambda_{T_1} = \lambda_{S_1}$, $\lambda_{T_2} = -\lambda_{S_2}$, and $\lambda_{T_3} = \lambda_{S_3}$, thus we can argue similarly as in case *(i)*.

(iii) Linearization around the equilibrium E_Q yields the three eigenvalues $\lambda_{Q_1} = B(\mu-\beta_Q)/(b+d) + (b+d+\mu)\beta_T/\beta_Q - \theta$ and

$$\lambda_{Q_{2,3}} = \frac{b(\mu-\beta_Q) \pm \sqrt{b(4\mu(b+d)^2 + \beta_Q\left(-4(b+d)^2 - 2b\mu + b\beta_Q\right) + 4(b+d)^3 + b\mu^2)}}{2(b+d)}.$$

$R_2 > 1$ is needed for the existence of E_Q. If we add the terms in λ_{Q_1}, it is easy to see that the numerator of the fraction is the difference of the numerator and the denominator of the reproduction number R_3, which means that it is negative if and only if $R_3 < 1$. The absolute value of the term under the square root in the nominator of λ_{Q_2}, resp. λ_{Q_3} is less than that of the first term which itself is negative as $\beta_Q > \mu$ follows from $R_2 > 1$. Thus, the last two eigenvalues always have negative real parts if $R_2 > 1$.

(iv) Linearizing around E_{QT}, we get the eigenvalues $\lambda_{QT_1} = -\lambda_{Q_1}$, $\lambda_{QT_2} = \lambda_{Q_2}$, and $\lambda_{QT_3} = \lambda_{Q_3}$, from which the assertion follows. □

2 Persistence and Global Stability

We shall use some notions and theorems from [3].

Definition 1 Let X be a nonempty set and $\rho : X \to \mathbb{R}_+$. A semiflow $\phi : \mathbb{R}_+ \times X \to X$ is called *uniformly weakly ρ-persistent*, if there exists some $\varepsilon > 0$ such that

$$\limsup_{t\to\infty} \rho(\Phi(t,x)) > \varepsilon \qquad \forall x \in X,\ \rho(x) > 0.$$

Φ is called *uniformly (strongly) ρ-persistent* if there exists some $\varepsilon > 0$ such that

$$\liminf_{t\to\infty} \rho(\Phi(t,x)) > \varepsilon \qquad \forall x \in X,\ \rho(x) > 0.$$

A set $M \subseteq X$ is called *weakly ρ-repelling* if there is no $x \in X$ such that $\rho(x) > 0$ and $\Phi(t,x) \to M$ as $t \to \infty$.

System (1) generates a continuous flow on the state space $X := \{(S, T, Q) \in \mathbb{R}^3_+\}$.

Theorem 1 *$S(t)$ is always uniformly persistent. $T(t)$ is uniformly persistent if $R_1 > 1$ and $R_2 < 1$ as well as if $R_2 > 1$ and $R_3 > 1$. $Q(t)$ is uniformly persistent if $R_2 > 1$.*

Proof The proof of the first assertion can be performed similarly as in [2, Theorem 4.3]. To prove the assertions about the persistence of $T(t)$ and $Q(t)$, we need some further theory from [3].

For the state of the system, we will use the notation $x = (S, T, Q) \in X$. We define the ω-limit set of a point $x \in X$ as usual by

$$\omega(x) := \{y \in X : \exists\{t_n\}_{n\geq 1} \text{ such that } \ t_n \to \infty, \Phi(t_n, x) \to y \text{ as } n \to \infty\}.$$

Let $\rho(x) = T$. Consider the invariant extinction space $X_T := \{x \in X : \rho(x) = 0\} = \{(S, 0, Q) \in \mathbb{R}^3_+\}$. The case $R_1 > 1$ and $R_2 < 1$ can be handled exactly as in [2, Theorem 4.3].

Let us now suppose that $R_2 > 1$ and $R_3 > 1$ hold. Following [3, Chap. 8], we examine the set $\Omega := \cup_{x\in X_T}\omega(x)$ for which in this case we have $\Omega = \left\{E_S, E_Q\right\}$. First we show weak ρ-persistence. To apply Theorem 8.17 of [3], we let $M_1 = \{E_S\}$ and $M_2 = \left\{E_Q\right\}$. We have $\Omega \subset M_1 \cup M_2$ and $\{M_1, M_2\}$ is acyclic and M_1 and M_2 are isolated, invariant and compact. We have to show that M_1 and M_2 are weakly ρ-repelling, then by [3, Chap. 8], the weak persistence follows.

Let us first assume that M_1 is not weakly ρ-repelling, i.e., there exists a solution with $\lim_{t\to\infty} (S(t), T(t), Q(t)) = (1, 0, 0)$ such that $T(t) > 0$. By $R_2 > 1$ and $R_3 > 1$,

$$R_2 R_3 = \frac{(b + d)\beta_T}{d\theta + b(\beta_Q + \theta - \mu)} > 1,$$

i.e., $\beta_T > \theta + (\beta_Q - \mu)b/(b + d)$. For t large enough we have $S(t) > 1 - \varepsilon$ and $Q(t) < \varepsilon$, so we can give the following estimation for $T(t)$:

$$\begin{aligned} T'(t) &= T(t)(\beta_T S(t) - \beta_Q Q(t) - \theta - b) > T(t)(\beta_T - \beta_T\varepsilon - \beta_Q\varepsilon - \theta - b) \\ &> T(t)\left(\frac{b}{b + d}(\beta_Q - \mu) - \varepsilon(\beta_T + \beta_Q) - b\right) \\ &= T(t)\left(\frac{b}{b + d}(\beta_Q - \mu - b - d) - \varepsilon(\beta_T + \beta_Q)\right), \end{aligned}$$

which is positive for ε small enough, since $R_2 > 1$ implies $\beta_Q > \mu + b + d$, thus $T(t) \to 0$ cannot hold.

Now we assume that M_2 is not weakly ρ-repelling, thus, there exists a solution with $\lim_{t\to\infty} (S(t), T(t), Q(t)) = (b+d+\mu)/\beta_Q, 0, b(\beta_Q - b - d - \mu)/(\beta_Q(b+d))$ and $T(t) > 0$. For any ε, for t large enough we can give the following estimations for $T'(t)$:

$$\begin{aligned} T'(t) &= T(t)(\beta_T S(t) - \beta_Q Q(t) - \theta - b) \\ &> T(t)\left(\beta_T\left(\frac{b + d + \mu}{\beta_Q} - \varepsilon\right) - \beta_Q\left(\frac{b(\beta_Q - b - d - \mu)}{(b + d)\beta_Q} + \varepsilon\right) - \theta - b\right) \end{aligned}$$

$$= T(t)\left(\frac{\beta_T(b+d+\mu)}{\beta_Q} - \frac{b(\beta_Q - b - d - \mu)}{b+d} - \theta - b - \varepsilon(\beta_T + \beta_Q)\right),$$

which is positive for ε small enough, since $R_3 > 1$.

The persistence of $Q(t)$ for $R_2 > 1$ can be proved using the same methods. The steps are analogous to those of the corresponding part of [2, Theorem 4.3] with only a slight modification needed. □

Using our theorem about persistence, in this section, we show that our LAS results extend to global asymptotic stability (GAS) results.

Theorem 2 *Equilibrium E_S is GAS if $R_1 \leq 1$ and $R_2 \leq 1$.*

Proof The proof is analogous to that of [2, Theorem 5.1] □

Theorem 3 *Equilibrium E_T is GAS stable on $X \setminus X_T$ if $R_1 > 1$ and $R_2 \leq 1$. On X_T, E_S is globally asymptotically stable.*

Proof The proof is analogous to that of [2, Theorem 5.2] □

Theorem 4 *Let us suppose $R_2 > 1$. Then the following statements hold:*

(i) If $R_3 \leq 1$ and $R_1 \leq 1$, then E_Q is GAS on $X \setminus X_Q$ and E_S is GAS on X_Q where $X_Q := \{x \in X : \{(S, T, 0) \in \mathbb{R}^3_+\}$, i.e., the extinction space of Q.
(ii) If $R_3 \leq 1$ and $R_1 > 1$, then E_Q is GAS on $X \setminus X_Q$ and E_T is GAS on X_Q.
(iii) If $R_3 > 1$, then E_{QT} is GAS on $X \setminus (X_Q \cup X_T)$, E_T is GAS on X_Q, E_Q is GAS on X_T.

Proof Let us introduce the notation $F(t) := S(t) + T(t)$. With this notation, we can transcribe system (1) to the two-dimensional system

$$\begin{aligned} F'(t) &= -\beta_Q F(t)Q(t) + \mu Q(t) + b - bF(t), \\ Q'(t) &= \beta_Q F(t)Q(t) - \mu Q(t) - bQ(t) - dQ(t). \end{aligned} \tag{4}$$

This system has the two positive equilibria $(1, 0)$ and

$$(F^*, Q^*) := \left(\frac{b+d+\mu}{\beta_Q}, \frac{b(\beta_Q - b - d - \mu)}{\beta_Q(b+d)}\right).$$

To show that the limit of each solution of this system is one of these two equilibria, according to the Poincaré–Bendixson theorem, all we have to prove is that system (4) does not have any periodic solutions. To show this, we apply Dulac's criterion using the Dulac function $D(Q, J) = 1/Q$. We have

$$\frac{\partial}{\partial F}\frac{-\beta_Q QF + \mu Q + b - bF}{Q} + \frac{\partial}{\partial Q}\frac{-bQ - dQ + \beta_Q FQ - \mu Q}{Q} = -\frac{b + Q\beta_Q}{Q} < 0.$$

From the previous section, we know that $Q(t)$ is persistent for $R_2 > 1$; thus, the limit of each solution started in $X \setminus X_Q$ is a subset of the set $\{x \in X : \{(S, T, Q^*) \in$

$\mathbb{R}^3_+ : S + T = F^*\}$. Thus, on the limit set the equation for $T(t)$ takes the form

$$T'(t) = \beta_T(F^* - T(t))T(t) - \beta_Q T(t)Q^* - \theta T(t) - bT(t) = -\beta_T T^2(t) + \gamma T(t),$$

where $\gamma = \beta_T F^* - \beta_Q Q^* - \theta - b$. The solution started from $T(0) = 0$ is the function $T(t) \equiv 0$. The nontrivial solutions of this logistic equation are $T(t) = \gamma C e^{\gamma t}/\beta_T C e^{\gamma t} + 1$ for $C \in \mathbb{R}_+$. It is easy to see that $\gamma > 0$ if and only if $R_3 > 1$. Thus, for $R_3 \leq 1$, $\lim_{t\to\infty} T(t) = 0$ and the limit of solutions started in $X \setminus X_Q$ is E_Q.

In the case $R_3 > 1$, we have

$$\lim_{t\to\infty} T(t) = \frac{\gamma}{\beta_T} = \frac{b+d+\mu}{\beta_Q} - \frac{\theta(b+d) - b\mu + b\beta_Q}{(b+d)\beta_T},$$

thus we obtain that the limit of solutions started in $X \setminus (X_T \cup X_Q)$ is E_{QT}. Solutions started in X_T tend to E_Q.

The limit set of solutions of Eq. (4) started in X_Q is the equilibrium $(1, 0)$. Thus, in this case, the equation for $T(t)$ on the limit set has the form $T'(t) = -\beta_T T^2(t) + \delta T(t)$ with $\delta = \beta_T - (\theta + b)$. Similarly to the previous case, the nontrivial solutions of this equation have the form $T(t) = \delta C e^{\delta t}/\beta_T C e^{\delta t} + 1$ for $C \in \mathbb{R}_+$. We have $\delta > 0$ if and only if $R_1 > 1$. Thus, for $R_1 \leq 1$, $T(t) \to 0$ $(t \to \infty)$ and the limit of solutions started in X_Q is E_S, while for $R_1 > 1$ we obtain $\lim_{t\to\infty} T(t) = \delta/\beta_T = 1 - (\theta + b)/\beta_T$, i.e., solutions started in X_Q tend to E_T. To complete the proof of the theorem, we notice that $R_2 > 1$ and $R_3 > 1$ imply $R_1 > 1$:

$$1 < R_2 R_3 = \frac{\beta_T}{b+\theta} \frac{(b+d)(b+\theta)}{d\theta + b(\beta_Q + \theta - \mu)} = R_1 \frac{b^2 + db + b\theta + d\theta}{d\theta + b\theta + b\beta_Q - b\mu} < R_1.$$

□

Finally, we comment on the impact of the disease-induced mortality d. For $d = 0$ we retrieve the results of [2]. Increasing d first decreases the total population without changing the qualitative dynamics. Sufficiently large d drives R_2 below 1. In this case, the disease dies out and the persistence of the parasites is determined by R_1.

Acknowledgement Research supported by ERC Starting Grant Nr. 259559 and by the European Union and cofunded by the European Social Fund under the projects TÁMOP-4.2.2.A-11/1/KONV-2012-0073 “Telemedicine-focused research activities on the field of Mathematics, Informatics and Medical sciences” and TÁMOP 4.2.4. A/2-11-1-2012-0001 “National Excellence Program—elaborating and operating an inland student and researcher personal support system.”

References

1. Brouqui, P., Raoult, D.: Arthropod-borne diseases in homeless. Ann. N. Y. Acad. Sci. **1078**, 223–235 (2006)
2. Dénes, A., Röst, G.: Global dynamics for the spread of ectoparasite-borne diseases. Nonlinear Anal. Real World Appl. **18**, 100–107 (2014)
3. Smith, H.L., Thieme, H.R.: Dynamical Systems and Population Persistence (Graduate Studies in Mathematics, Vol. 118). AMS, Providence (2011)

Temperature Induced Cubic-to-Tetragonal Transformations in Shape Memory Alloys Using a Phase-Field Model

R. Dhote, H. Gomez, R. Melnik and J. Zu

Abstract Shape memory alloys (SMAs) exhibit hysteresis behaviors upon stress- and temperature-induced loadings. In this chapter, we focus on numerical simulations of microstructure evolution of cubic-to-tetragonal martensitic phase transformations in SMAs in 3D settings under the dynamic loading conditions. A phase-field (PF) model has been developed to capture coupled dynamic thermo-mechanical behavior of such SMA structures and the system of governing equations have been solved numerically using the isogeometric analysis. Temperature induced reverse and forward transformations have been applied to a cubic SMA specimen, starting with evolved accommodated martensitic microstructure. We have observed that during the forward transformation, the martensitic variants nucleate abruptly. The transient microstructures are aligned along [110] planes, which is in accordance with the crystallographic theory and experimental results.

1 Introduction

Shape memory alloys (SMAs) have been widely used in commercial applications and studied in the research community for their interesting shape recovering, hysteretic properties, and complex microstructure morphology [1–3]. Most of these

R. Dhote (✉)
Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: rakesh.dhote@mail.utoronto.ca

R. Dhote · J. Zu
University of Toronto, Toronto, ON, Canada

H. Gomez
University of A Coruña, Coruña, Spain
e-mail: hgomez@udc.es

R. Melnik
The MS2Discovery Interdisciplinary Research Institute,
Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: rmelnik@wlu.ca

J. Zu
e-mail: zu@mie.utoronto.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_26

studies/applications have been developed to model/utilize static or quasistatic behaviors of SMAs. There exists a number of areas (e.g., energy absorption and vibration damping, to just name a few) where the dynamic behavior of SMAs is essential. Our better understanding of microstructure evolution and its effect on SMA properties, upon dynamic loading, will help in the development of better models and devices.

In this contribution, we present a 3D model to study cubic-to-tetragonal phase transformations in SMAs. The model is developed based on a phase-field approach and the phenomenological Ginzburg–Landau theory [4–6]. A Ginzburg–Landau free energy functional is defined in terms of two (deviatoric) strain based order parameters, whose roots define a phase in a system at a particular temperature. The austenite phase is represented by a cubic arrangement of atoms which occur at higher temperatures. The tetragonal arrangement of atoms occur at lower temperatures resulting in martensitic variants, which are energetically equivalent. The governing equations of the mathematical model are derived from the conservation laws of mass, momentum, and energy [7]. The developed model has highly nonlinear hysteretic behavior, bidirectional thermomechanical coupling and higher (fourth) order spatial differential terms [6]. The fourth-order differential terms define a smoothly varying diffuse interface between austenite and martensite variants or between martensite variants. Traditionally, such higher-order differential models have been numerically solved using a finite difference, spectral methods, etc. [5]. These methods have known limitations in terms of geometric flexibility of a domain. An isogeometric analysis (IGA) is a geometrically flexible method that can be used to study real-world devices of complex shape. IGA offers advantages in exact geometric representations, higher-order continuity, accuracy, and robustness [8]. In [6], we first reported the use of IGA methodology to study microstructure evolution for the 3D cubic-to-tetragonal phase transformations in SMAs. In this chapter, we study microstructure evolution in SMAs under temperature-induced transformations.

In Sect. 2, we present the phase-field model describing the cubic-to-tetragonal transformations in SMAs and its numerical implementation based on the IGA. In Sect. 3, we study microstructure evolution in a SMA domain under reverse and forward transformations starting with accommodated twinned microstructures. Conclusions are discussed in Sect. 4.

2 Mathematical Model and Numerical Implementation

The following three-well Ginzburg-Landau free energy functional can be used to describe cubic-to-tetragonal phase transformations in SMAs. The functional is expressed in terms of strain-based order parameters and temperature, as

$$\begin{aligned}\mathcal{F} =& \frac{a_{31}}{2}\left[e_1^2\right] + \frac{a_{36}}{2}\left[e_4^2+e_5^2+e_6^2\right] + \frac{a_{32}}{2}\tau(e_2^2+e_3^2) + \frac{a_{33}}{2}e_3(e_3^2-3e_2^2)\\ &+ \frac{a_{34}}{2}(e_2^2+e_3^2)^2 + \frac{k_g}{2}\left[(\nabla e_2)^2+(\nabla e_3)^2\right],\end{aligned} \tag{1}$$

where a_{ij}, k_g are the material parameters and τ is the temperature coefficient [4–6]. The strain e_1 represents bulk strain, e_2 and e_3 represent deviatoric strains, and e_4, e_5, and e_6 represent shear strains. The e_i strains are defined using the Cauchy–Lagrange strain tensor as $e_{ij} = \left[\left(\partial u_i/\partial x_j\right) + \left(\partial u_j/\partial x_i\right)\right]/2$ (using the repeated index convention), where $\mathbf{u} = \{u_i\}|_{i=1,2,3}$ are the displacements along x, y, and z directions, respectively. The first and second terms in the functional represent bulk and shear energy, respectively. The next three terms represent the Landau energy that define phase transformations between austenite and martensites and between martensite variants. The last term represents the gradient energy that describes nonlocal elastic behavior. The Landau energy has three minima having equal energies, corresponding to the three martensitic variants, below the critical temperature, one minima, corresponding to the austenite phase, above the critical temperature. The system has degenerate state near the critical temperature.

The mathematical model is described by conservation laws of mass, momentum, and energy [6, 7] as

$$\dot{\mathbf{u}} = \mathbf{v}, \tag{2}$$

$$\rho\dot{\mathbf{v}} = \nabla \cdot \sigma + \nabla \cdot \sigma' + \mathbf{f}, \tag{3}$$

$$\rho\dot{e} - \sigma^T : (\nabla v) + \nabla \cdot \mathbf{q} = g, \tag{4}$$

where ρ is the mass density, q is the Fourier heat flux vector, f and g are external mechanical and thermal loadings. The stress tensors σ and dissipation stress tensors σ' are defined as

$$\boldsymbol{\sigma} = \frac{\partial \mathcal{F}}{\partial e_{ij}}, \qquad \boldsymbol{\sigma}' = \frac{\partial \mathcal{R}}{\partial \dot{e}_{ij}}. \tag{5}$$

The Rayleigh dissipation energy functional $\mathcal{R} = \eta/2 \sum \dot{e}_i^2$ is added to stabilize the microstructure quickly, where η is the dissipation coefficient.

The developed model has highly nonlinear hysteretic behavior, thermomechanical coupling, and fourth-order spatial differential terms. The weak form of the governing Eqs. (2)–(4) are obtained by multiplying them with weighting functions and transforming them by using the integration by parts. We implement the weak form of the governing equations in the IGA for numerical solution. The semidiscrete formulation, where the space is discretized using the Galerkin finite element scheme and time is treated as continuous has been described in [6].

3 Numerical Simulations

The simulations in this section are conducted on a cubic domain with 80 nm side. All the simulations have been performed on the Sharcnet clusters utilizing 64 processors (4 processors each in three directions) with 1 GB memory each. The decomposed domain, in each processor, is discretized with 16 quadratic $\mathcal{C}^1$-continuous nonuniform rational basis spline (NURBS) basis in each direction. The periodic boundary

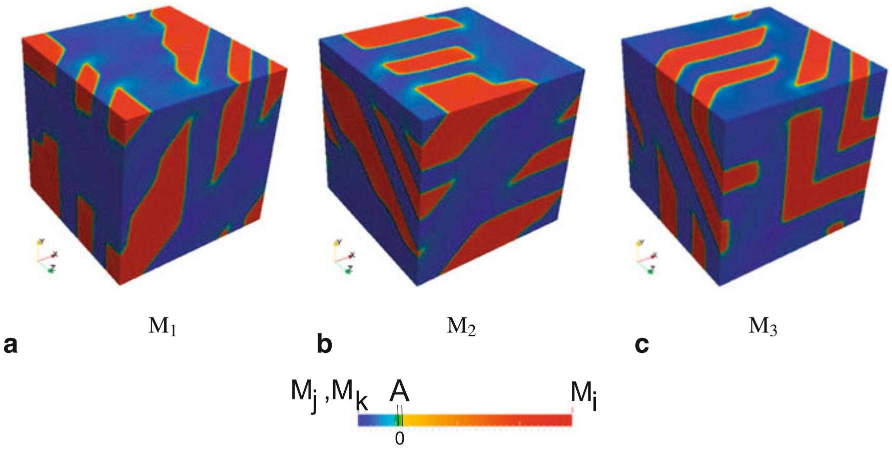


Fig. 1 (*Color online*) Self accommodated microstructure in a cube domain with **a** M$_1$, **b** M$_2$, **c** M$_3$ martensitic variants (*red color* represent M$_i$ variant, *blue* represent the remaining two variants M$_j$ and M$_k$, and *green color* represents austenite (A) phase)

conditions have been used in the structural physics and insulated for the thermal physics. The material parameters are identical to those used in [5]. The simulations have been carried out to study microstructure evolution under temperature induced reverse and forward phase transformations, without application of a mechanical load.

We first obtain the accommodated twinned microstructure in a domain by allowing the system to evolve, starting with initial random conditions in displacement u and temperature coefficient $\tau = -1.2$. The system minimizes its energy and stabilizes into accommodated twinned martensitic variants. Figure 1 shows the three variants of martensites M$_1$, M$_2$, and M$_3$ corresponding to martensite phase (tetragonal) aligned along the x, y, and z directions, respectively. The microstructures are characterized by the axial strain values (e.g., martensite M$_1$ is represented by $\epsilon_{11} > 0$, i.e., tetragonal variant elongated along the x-direction). The red color in each subplot of Fig. 1 represents M$_i$ variant and blue represents the remaining two variants, as shown in the color spectrum in the figures. The competition between bulk, shear, and gradient energy results in three variants accommodated in a herringbone structure with domain walls aligned along [110] planes, which is in accordance with the crystallographic theory and experimental results [1, 9].

Next, we perform a temperature-induced reverse transformation (RT, martensite → austenite) starting with the evolved microstructure in the previous step. The thermal loading is applied on a domain with $\bar{g} = 0.05\bar{t}$ in the dimensionless units (bar shows the dimensionless variable). Figure 2 shows the time snapshot of the microstructure at intermediate time (first row) and at the end of unloading (second row). The domain walls no longer remain distinct and sharp, as compared to Fig. 1, and extinct at the end of thermal loading. This observation is in accordance with the experimental evidence [3].

Finally, we use the evolved austenite microstructure at the end of loading cycle as the initial condition to the forward transformation (FT, austenite → martensite)

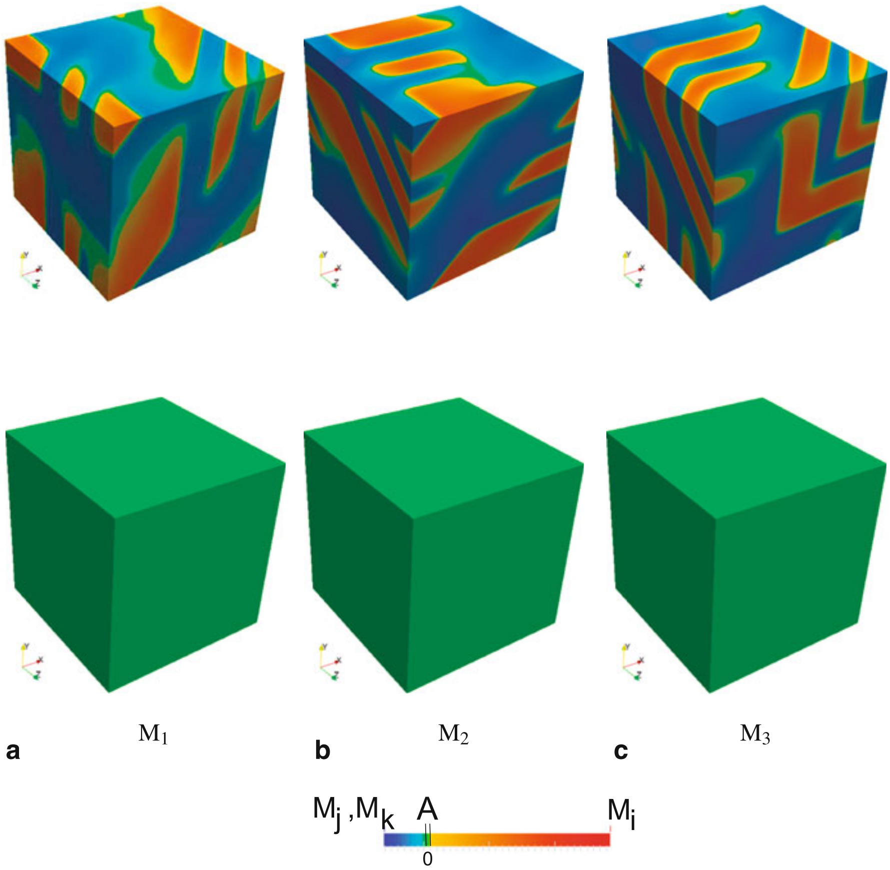


Fig. 2 (*Color online*) Microstructure evolution during RT: the first row shows the microstructure at $\bar{t} = 950$ and the second row at the end of loading cycle $\bar{t} = 1080$ (*red color* represents M_i variant, *blue* represents the remaining two variants M_j and M_k, and *green color* represents austenite (A) phase). *RT* reverse transformation

by applying the thermal loading on a domain with $\bar{g} = -0.1\bar{t}$ in the dimensionless units. The martensitic microstructure evolve abruptly at $\tau \approx -5$ at approximately 1500 time units during unloading. The transient martensitic variants on nucleation are shown in Fig. 3.

The average temperature coefficient τ evolution in SMA domain during microstructure evolution, RT, and FT are shown in Fig.4. The nucleation of martensitic variant from austenite during FT is seen with a jump in τ at approximately 1500 time units.

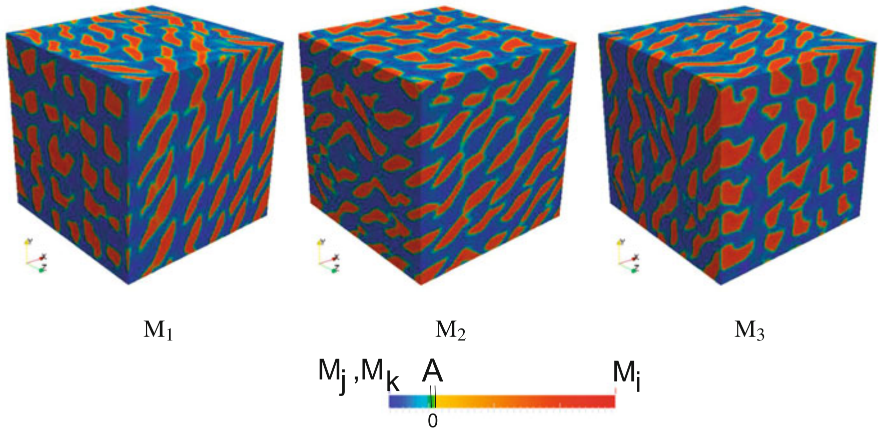


Fig. 3 (*Color online*) Transient microstructure at $\bar{t} \approx 1500$ during the FT (*red color* represents M_i variant, *blue* represents the remaining two variants M_j and M_k, and *green color* represents austenite (A) phase). *FT* forward transformation

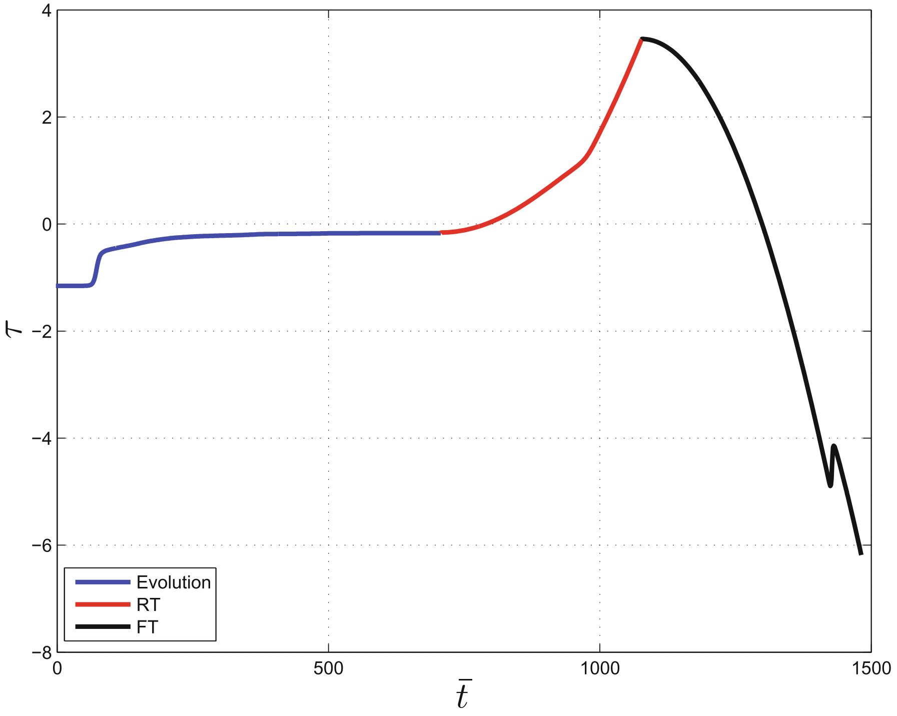


Fig. 4 (*Color online*) Average temperature coefficient τ plot during microstructure evolution (*blue*), RT (*red*), and FT (*black*)

4 Conclusions

The fully coupled thermomechanical model to describe cubic-to-tetragonal phase transformations in SMAs has been developed and numerically implemented in the IGA.

We have numerically analyzed the temperature induced reverse and forward phase transformations in SMAs. It has been found that the domain walls between martensite variants are aligned in accordance with the crystallographic theory and experimental results. We have also captured the abrupt nucleation of martensitic variants during the reverse transformation.

Acknowledgments This work was made possible with the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada. Support of NSERC and CRC program is also gratefully acknowledged.

References

1. Bhattacharya, K.: Microstructure of Martensite: why it forms and how it gives rise to the shape-memory effect. Oxford University Press, ISBN: 9780198509349 (2003)
2. Otsuka, K., Wayman, C.: Shape Memory Materials. Cambridge University Press, ISBN: 052144487 (1999)
3. Lagoudas, D.: Shape Memory Alloys: Modeling and Engineering Applications. Springer, ISBN: 9780387476858 (2008)
4. Barsch, G., Krumhansl, J.: Twin boundaries in ferroelastic media without interface dislocations. Phys. Rev. Lett. **53**(11), 1069–1072 (1984)
5. Ahluwalia, R., Lookman, T., Saxena, A.: Dynamic strain loading of cubic to tetragonal martensites. Acta Materialia **54**(8), 2109–2120 (2006)
6. Dhote, R., Gomez, H., Melnik, R., Zu, J.: Isogeometric analysis of coupled thermo-mechanical phase-field models for shape memory alloys using distributed computing. Procedia Comput. Sci. **18**, 1068–1076 (2013)
7. Melnik, R., Roberts, A., Thomas, K.: Computing dynamics of copper-based SMA via centre manifold reduction of 3D models. Comput. Mater. Sci. **18**(3), 255–268 (2002)
8. Cottrell, J., Hughes, T., Bazilevs, Y.: Isogeometric Analysis: Toward Integration of CAD and FEA. John Wiley & Sons, ISBN: 9780470748732 (2009)
9. Sapriel, J.: Domain-wall orientations in ferroelastics. Phys. Rev. B **12**(11), 5128 (1975)

A Study of Brain Biomechanics Using Hamilton's Principle: Application to Hydrocephalus

Corina S. Drapaca and Justin A. Kauffman

Abstract Hydrocephalus is a serious neurological disorder characterized by abnormalities in the circulation of cerebrospinal fluid (CSF) within the brain. Unfortunately, the response of the patients who have been treated for hydrocephalus continues to be poor and thus better therapy protocols are desperately needed. Mathematical models of CSF dynamics and CSF–brain interactions could play important roles in the design of improved, patient-specific treatments. To capture some of brain's dynamics during the evolution of hydrocephalus we propose a new mathematical model using Hamilton's principle. We assume the existence of current healthy healing and abnormal inflammation states and investigate the relationship between these states using volumetric data of healthy and untreated hydrocephalic mice.

1 Introduction

Hydrocephalus is a brain disease caused by abnormalities in the cerebrospinal fluid (CSF) circulation resulting in ventricular dilation, brain compression, and in some cases an increase in the intracranial pressure. The treatment is based on CSF flow diversion and continues to suffer from poor outcomes [8]. Therefore, there is an urgent need to design better therapy protocols for hydrocephalus. An important step in this direction is the development of predictive mathematical models that better explain the fundamental science behind hydrocephalus. While modern models that focus on how best to relate brain's mechanics to its biochemistry are essential in enhancing our understanding of mechanisms of hydrocephalus (and brain biomechanics in general), the lack of experimental data needed by these models as well as the complexity of the corresponding computations make these models difficult to use in clinical applications for now. In this chapter, we propose a mathematical model using Hamilton's principle that captures some of the brain's dynamics during the

C. S. Drapaca (✉) · J. A. Kauffman
Pennsylvania State University, University Park, PA 16802, USA
e-mail: csd12@psu.edu

J. A. Kauffman
e-mail: jak5378@psu.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_27

evolution of hydrocephalus. We define current healthy healing and abnormal inflammation states and investigate the relationship between these states using volumetric data of healthy and untreated hydrocephalic mice reported in [3].

2 Mathematical Model

In this section, we present a one-dimensional model for brain dynamics using Hamilton's principle. For simplicity, we assume that in the healthy state only small macro-deformations (linear kinematics) can occur in the brain tissue and the mechanical response of the tissue at the macroscopic level is linear viscoelastic of Kelvin–Voight type. The one-dimensional brain tissue of length L has one fixed boundary ($x = 0$) at the interface with the meninges surrounding the brain and one moving boundary ($x = L$) at the interface with the ventricular CSF which undergoes macroscopic displacements caused by the heart pulsations and healthy aging. We assume further that there are two biological processes that influence brain's functionality:

1. A microstructural healthy healing of brain controlled by functional microglial cells [5] and we denote by $\psi_h(x,t)$ the current healing state function, and
2. A microstructural sustained inflammation of brain caused by some dysfunctional microglial cells [5] that progresses slowly throughout the entire life and we denote by $\psi_i(x,t)$ the current inflammation state function.

Our second assumption is based on clinical studies [5, 7] that have shown that prolonged inflammation plays an important role in the process of normal aging and neurodegeneration diseases. We suggest excessive inflammation of the choroid plexus (anatomical structure located at the interface between the brain tissue and the ventricular CSF which is involved in the CSF production) as one possible mechanism for the onset of postinfectious and posthemorrhagic pediatric hydrocephalus, as well as for the onset of normal pressure hydrocephalus in some older people.

In what follows, we generalize the theoretical concepts introduced in [1] and adapt them to our model's assumptions. We propose a Lagrangian of the form:

$$\mathcal{L} = \int_0^L \left[\frac{1}{2} m \dot{u}^2 + \frac{1}{2} m\alpha \left(\dot{\psi}_h^2 + \dot{\psi}_i^2 \right) - \frac{1}{2} \bar{E}(\psi_h, \psi_i) A u'^2 - \frac{1}{2} \beta \left({\psi'}_h^2 + {\psi'}_i^2 \right) \right] dx \tag{1}$$

where m is the mass density of the one-dimensional brain tissue of length L and constant cross-sectional area A, $u(x,t)$ is the macroscopic displacement, $\bar{E}$ is the effective macroscopic elastic modulus, α and β are positive constants. We notice that we work with a special Lagrangian, since, in general, the coefficients of the terms $\dot{\psi_h}^2$, $\dot{\psi}_i^2$, ${\psi'}_h^2$, ${\psi'}_i^2$ in Eq. (1) do not need to be the same. For simplicity, we denote by $\dot{u} = \frac{\partial u}{\partial t}$, $u' = \frac{\partial u}{\partial x}$. The second and fourth terms of Eq. (1) represent microstructural kinetics and, respectively, energies caused by the evolution of the brain's microstructure due to normal healing and prolonged inflammation.

As in [1], we define the virtual work done by nonconservative forces as:

$$\delta \mathcal{W} = \int_0^L \left[f\delta u - c\dot{u}'\delta u' - \Psi_h(\psi_h, \dot{\psi}_h, \psi_h')\delta\psi_h - \Psi_i(\psi_i, \dot{\psi}_i, \psi_i')\delta\psi_i \right] dx + F\delta u|_L \tag{2}$$

where f is a body force per unit length, $c\dot{u}'$ is the linear damping term of the Kelvin–Voight model with c the viscosity, and Ψ_h, Ψ_i are generalized forces that are work conjugates of the *evolution variables* ψ_h, ψ_i. We denote by F the concentrated load on the ventricular CSF–brain interface. In order for (2) to be thermodynamically consistent, we apply Clausius–Duhem inequality and obtain (for details see [2]):

$$c \geq 0,\ \Psi_h\dot{\psi}_h \geq 0,\ \Psi_i\dot{\psi}_i \geq 0 \tag{3}$$

The nonconservative form of Hamilton's principle: $\int_{t_1}^{t_2} (\delta\mathcal{L} + \delta\mathcal{W})dt = 0$, for the independent variables δu, $\delta\psi_h$, $\delta\psi_i$ that vanish at arbitrary times t_1, t_2, gives the following system of partial differential equations:

$$-m\ddot{u} + A\left(\bar{E}u'\right)' + f + c\dot{u}'' = 0 \tag{4}$$

$$-m\alpha\ddot{\psi}_h - \frac{1}{2}Au'^2\frac{\partial\bar{E}}{\partial\psi_h} + \beta\psi_h'' - \Psi_h = 0 \tag{5}$$

$$-m\alpha\ddot{\psi}_i - \frac{1}{2}Au'^2\frac{\partial\bar{E}}{\partial\psi_i} + \beta\psi_i'' - \Psi_i = 0. \tag{6}$$

To system (4)–(6) we add initial conditions, Dirichlet and/or Neumann boundary conditions at $x = 0,\ L$ for ψ_h, ψ_i, and $u(0,t) = 0,\ A\bar{E}u'(L,t) + c\dot{u}'(L,t) = F$. We start our analysis in the following simpler case: $\Psi_h = \Psi_i = F = f = 0$, $\psi_h = \psi_h(t)$, $\psi_i = \psi_i(t)$. We assume that the brain tissue becomes stiffer during the healthy healing period but slowly softer due to aging [6] (the softening of the tissue might facilitate abnormal inflammatory processes). Taking into account these facts and the approach proposed in [1] to study damage mechanics, we introduce the following expression for the dynamics of brain's effective elastic modulus $\bar{E}$: $\bar{E} = \lambda\psi_h(1-\psi_i)$ for $\psi_h \geq 1,\ 0 < \psi_i < 1$. The case $\psi_h = 1$ corresponds to no healing. With these simplifications, system (4)–(6) reduces to:

$$\ddot{\psi}_h = -\frac{\lambda}{2m\alpha}Au'^2(1-\psi_i),\ \ddot{\psi}_i = \frac{\lambda}{2m\alpha}Au'^2\psi_h \tag{7}$$

Our model has not been experimentally and/or clinically validated yet.

Table 1 Calculated Jacobians J from brain volumetric data [3]

Time (days)	J for healthy mice (J_{nm})	J for hydrocephalic mice (J_{hm})
18	1	1
22	1.032	1.090
23	1.039	1.268
28	1.071	1.280
85	1.250	1.424

3 Application to Hydrocephalus

To make some progress in understanding how inflammation could contribute to the onset and evolution of hydrocephalus using our model, we used brain volumetric data for healthy and untreated hydrocephalic mice published in [3]. For simplicity we took $\frac{\lambda}{2m\alpha} = 1$ in Eq. (7). Since we consider the one-dimensional case, the strain was calculated as follows:

$$u' = J - 1 \tag{8}$$

where the Jacobian of the deformation J is a measure of volume change during the deformation. Table 1 shows values of J calculated as the ratio of current brain volume over the initial brain volume.

We used the built-in Matlab function *polyfit* to find two quadratic fitting functions $J_{nm}(t),\ J_{hm}(t)$ for the data shown in Table 1 for healthy and, respectively, hydrocephalic mice. We replaced formula (8) into system (7) and obtained the following system of first-order linear ordinary differential equations:

$$\dot{\psi_h} = v_h,\ \dot{\psi_i} = v_i,\ \dot{v_h} = -(J-1)^2(1-\psi_i),\ \dot{v_i} = (J-1)^2\psi_h \tag{9}$$

The solution to system (9) must however satisfy the following constraints:

$$\psi_h \geq 1,\ 0 < \psi_i < 1,\ -\infty < v_h < \infty,\ -\infty < v_i < \infty \tag{10}$$

We denote by dt the step of a equally-spaced time discretization, and $[\psi_h^n, \psi_i^n, v_h^n, v_i^n]^T$ the solution corresponding to the discrete time point t_n. By applying an implicit scheme for the discretization of system (9), we obtained the following system of linear algebraic equations:

$$\begin{bmatrix} 1 & 0 & -dt & 0 \\ 0 & 1 & 0 & -dt \\ 0 & -dt(J(t_{n+1})-1)^2 & 1 & 0 \\ -dt(J(t_{n+1})-1)^2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \psi_h^{n+1} \\ \psi_i^{n+1} \\ v_h^{n+1} \\ v_i^{n+1} \end{bmatrix}$$

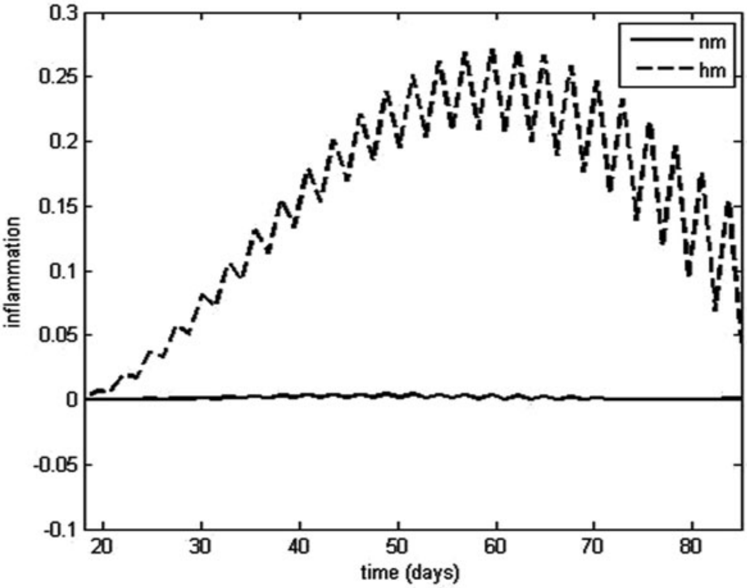


Fig. 1 Inflammation functions for healthy (nm) and hydrocephalic (hm) mice for the initial condition $[\psi_h^0, \psi_i^0, v_h^0, v_i^0] = [1, 0, 0, 0]$

$$= \begin{bmatrix} \psi_h^n \\ \psi_i^n \\ v_h^n - dt(J(t_{n+1}) - 1)^2 \\ v_i^n \end{bmatrix} \tag{11}$$

Problems (11) and (10) are a mixed complementarity problem and were solved separately for $J = J_{nm}$ and $J = J_{hm}$ using Matlab and the function *pathlcp* of the PATH solver [4] at each time step. We show results for two sets of initial conditions: (1) $[\psi_h^0, \psi_i^0, v_h^0, v_i^0] = [1, 0, 0, 0]$ and (2) $[\psi_h^0, \psi_i^0, v_h^0, v_i^0] = [1, 0, 2, 0]$. In the first case, the healing function remains constant equal to unity for healthy and hydrocephalic mice, while the inflammation increases slightly for healthy mice and a lot more for hydrocephalic mice (Fig. 1). For the second set of initial conditions where the initial healing speed is nonzero, the inflammation and healing functions increase much more for hydrocephalic mice than for healthy ones (Fig. 2). These results suggest that inflammation may be one possible mechanism for hydrocephalus. In addition, it appears that in hydrocephalus normal healing alone is not enough to reduce the excessive inflammation.

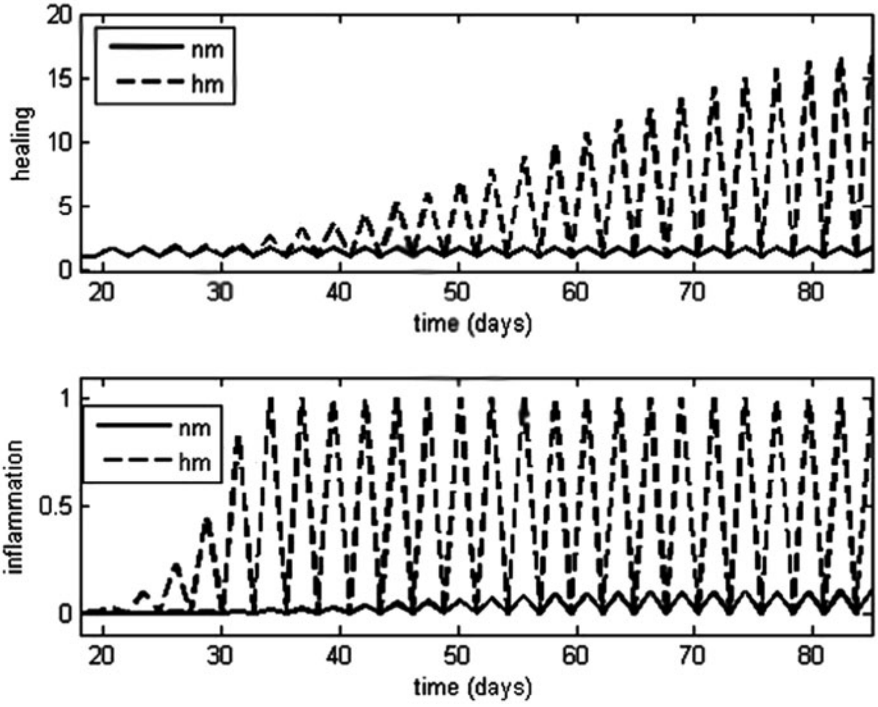


Fig. 2 Healing (*top*) and inflammation (*bottom*) functions for healthy (nm) and hydrocephalic (hm) mice for the initial condition $[\psi_h^0, \psi_i^0, v_h^0, v_i^0] = [1, 0, 2, 0]$

4 Conclusion

In this chapter, we proposed a mathematical model using Hamilton's principle to study hydrocephalus. We used volumetric data of healthy and untreated hydrocephalic mice to investigate possible relationships between healthy healing and prolonged inflammation states. Next we plan to validate our model and study different treatment effects.

Acknowledgments The first author would like to thank her colleagues Drs. Cusumano, Schiff, and Costanzo for insightful discussions and comments on this work.

References

1. Cusumano, J.P., Roy, A., Li, Q.: Damage dynamics, rate laws, and failure statistics via Hamilton's principle. Meccanica, **50**(1), 77–98 (2015)
2. Kauffman, J.: Mathematical models of brain and cerebrospinal fluid dynamics: application to hydrocephalus, Master's Thesis, Pennsylvania State University (2013)

3. Mandell, J., Neuberger, T., Drapaca, C., Webb, A., Schiff, S.: The dynamics of brain and cerebrospinal fluid growth in normal versus hydrocephalic mice. J. Neurosurg. Pediatr. **6,** 1–10 (2010)
4. Ferris, M.C., Munson, T.S. Interfaces to PATH 3.0: Design, Implementation and Usage. Computational Optimization and Applications **12**(1–3), 207–227 (1999)
5. Rivest, S.: Regulation of innate immune responses in the brain. Nat. Rev. Immunol. **9,** 429–439 (2009)
6. Sack, I., Beierbach, B., Wuerfel, J., Klatt, D., Hamhaber, U., Papazoglou, S., Martus, P., Braun, J.: The impact of aging and gender on brain viscoelasticity. Neuroimage **46**(3), 652–657 (2009)
7. Singh, T., Newman, A.B.: Inflammatory markers in population studies of aging. Ageing Res. Rev. **10**(3), 319–329 (2011)
8. Tuli, S., Alshail, E., Drake, J.: Third ventriculostomy versus cerebrospinal fluid shunt as a first procedure in pediatric hydrocephalus. Pediatr. Neurosurg. **30**(1), 11–15 (1999)

A Mathematical Model for Treatment Selection Literature

G. Duncan and W. W. Koczkodaj

Abstract Business intelligence (BI) tools and techniques, when applied to a data store of bibliographical references, can provide a researcher with valuable information and metrics. In contrast to specialized research platforms that provide a number of analysis tools, such as the Web of Knowledge^{TM} (WOK) or PubMed^{TM}, the techniques discussed in this chapter provide a more generalized approach that can be used with most bibliographical data sets as well as with a number of different analysis tools. As a point of reference, the system utilizes the WOK's Web of Science (WOS) database schema, chosen because it provides a comprehensive number of bibliographical information fields. This chapter will discuss how to transform WOK formatted data into an online analytical processing (OLAP) cube as well as provide a few examples of using this technology to analyze bibliographical information.

1 Introduction

Business intelligence (BI) "is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information" [1]. Coined in 1958 by IBM researcher Hans Peter Luhn, BI has since seen broad acceptance by the business world. BI provides a number of important applications in the enterprise, including measurement, analytics and reporting [1]. The core of BI is in utilizing large stores of data (referred to as data marts or data warehouses) to enable adhoc analysis, measurement metrics and data mining which can then be used in order to influence and guide business decisions. While BI has found a home in a number of business sectors such as in the financial and healthcare industries, using it to aid in the meta-analysis of reference information is virtually unknown. In general, BI can be applied to any relational data set.

G. Duncan (✉) · W. W. Koczkodaj
Computer Science, Laurentian University, 935 Ramsey Lake Road, Sudbury,
ON P3E 2C6, Canada
e-mail: gg_duncan@laurentian.ca

W. W. Koczkodaj
e-mail: wkoczkodaj@cs.laurentian.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_28

The sample implementation discussed within this chapter is based on Microsoft's standard information management and software development offerings. The database used will be SQL 2008R2 and standard TSQL, the front-end system is implemented in C# under Microsoft Visual Studio 2012 and analysis/OLAP services will be provided by Structured Query Language (SQL) Server Business Intelligence Development Studio 2008 deployed to an SQL Server Analysis Services server (SQL 2008R2). The choice of platform was made due to familiarity to the author.

Summary

The proposed system will be described in the following steps:

1. Extract and transform the bibliographical data such it conforms to the Web of Knowledge (WOK) format.
2. Serialize the transformed data into an SQL database.
3. Transform the serialized SQL data into data dimensions and a fact table arranged in a star schema.
4. Define the OLAP cube's data dimensions' measures and their inter-relationships.
5. Create an OLAP cube from the data and deploy the cube to an analysis server.

Once the data is deployed, standard BI and OLAP tools (such as Microsoft Excel, Microsoft SQL Server Business Intelligence Development Studio or Tableau) can be used in order to analyze the cube. Manual analysis of the cube via the use of Data Mining Extensions (DMX) or Multidimensional Expressions (MDX) will be considered out of scope in this document.

2 Extract and Transform

The WOK allows the user to download search results into a variety of formats, as does PubMed. Once downloaded, this data may need to be transformed into an intermediate format so that it's compatible with the underlying database schema (or at least compatible with the tools being used to import the data into the database). In order to perform such a transform of the data, the example implementation utilizes Extensible Markup Language (XML), a standards-based markup language that defines rules for encoding information in a human and machine readable format [4], and Extensible Stylesheet Language Transformations (XSLT) a language for transforming XML documents [5]. XML is an extremely well-supported format; it is the format of choice for many Internet data interchange protocols (see, for instance, [3]), which makes it particularly suitable as an intermediary format for extract and transform operations. See Fig. 1 for an overview of the process.

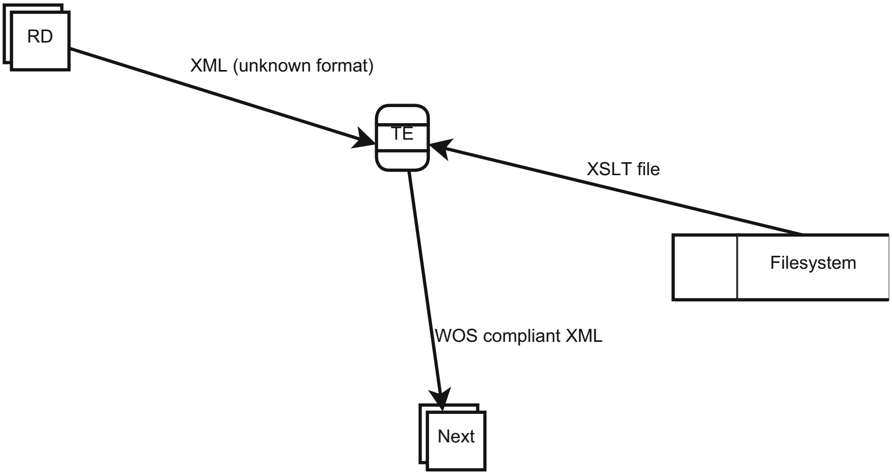
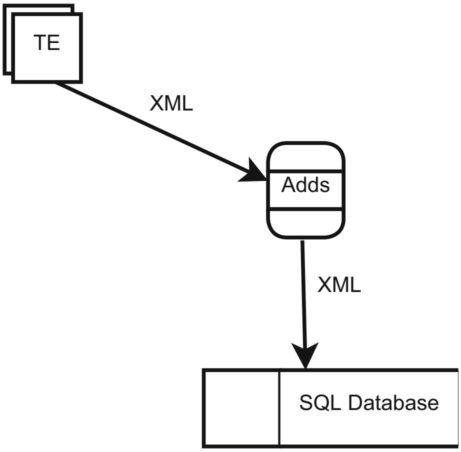


Fig. 1 Extract and transform Data Flow Diagram (DFD). The user downloads Extensible Markup Language (XML) from the reference database (RD) and sends it to the transform engine (TE). The appropriate Extensible Stylesheet Language Transformations (XSLT) files is loaded from the file system, and the transform is applied, creating an XML document compliant with the desired schema/format

Fig. 2 Serializing the transformed data to a database. The transform engine (TE) sends the Extensible Markup Language (XML) to the process that issues Transact-SQL (TSQL) commands to insert the information into the SQL database



3 Serializing to a Database

Once the data is in the necessary format, the next step is to serialize it to a database. In the case of the example implementation, the data is placed into an SQL database table that mimics the schema of it’s input data. This database will form the basis of the data warehouse from which the analysis services will query. It’s not necessary that this database be normalized. See Fig. 3 for a description of the process.

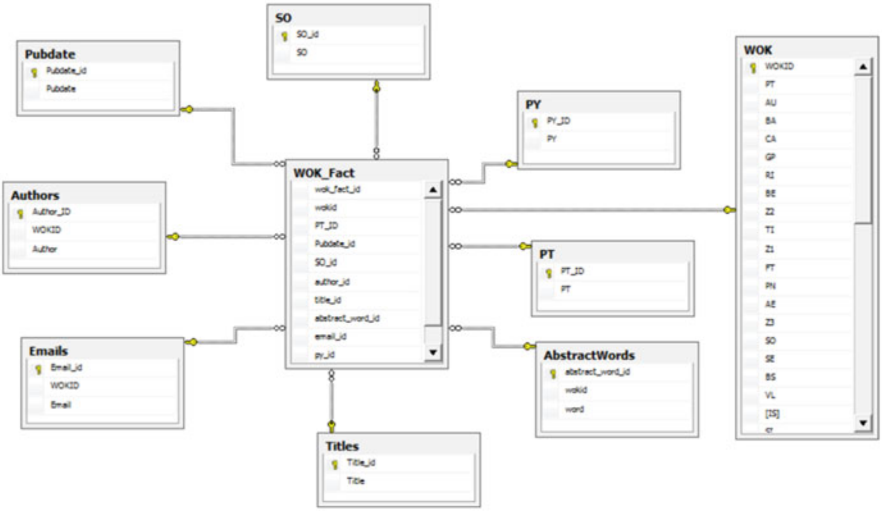


Fig. 3 Example implementation fact and dimension tables (data warehouse). Notice the "star" pattern of the layout

4 Data Dimensions, Fact Table, and the Star Schema

In order to create the OLAP cube, the data must be cut into dimensions and one or more fact tables must be created. The collection of dimensions, fact tables, and data tables is collectively called the "data warehouse." A data dimension is a "data set composed of individual, non-overlapping data elements," the primary purpose of which is to provide "filtering, grouping and labeling" [2]. A fact table is a table that joins all the relationships between the various dimensions. The exact methods used to create the dimension and fact tables will vary depending on the a number of factors, such as the granularity of the fact tables and dimensions and the number and types of dimensions.

It is important to consider the data type and content, since some may not be very appropriate as dimensions. In some cases, the data may have to be broken up or otherwise manipulated in order to enable sufficient levels of granularity for effective analysis. For example, in Web of Science (WOS) formatted records, author and email information are each stored as semicolon separated lists ("email1@domain.com; email2@domain.com"). In order for this field to be used as a data dimension, it is first necessary to separate out the individual values from their collated representations, and then individually insert them into the dimensional table. It is also important to consider how to key the dimensional information, (particularly the primary key, as this key will define one or more columns of the fact table). The typical technique is used to define the table's primary key as an integer "identity" column. This provides for each row to be unique by definition. Other methods include the use of multipart keys or object guides. Recall that the definition of a data dimension specifies that the data contained therein be nonoverlapping and thus distinct.

Once the data dimensions have been defined and created, the next step is to create one or more fact tables from this data. The fact table provides one row for each valid combination of the dimension tables' values.

The fact table is composed of all the foreign keys from each of the dimensional tables, along with any nondimensional data from the basis table.

Figure 3 presents the entity relationship diagram created for the example implementation, notice it's arrangement in the "star" pattern, where all dimensional tables are arranged around the central fact table.

4.1 Fact and Dimension Table Choices in the Example Implementation

For the example implementation, the following WOK record columns were chosen:

AU	Authors information
TI	Title
SO	Full source title
PD	Publication date
PY	Publication year
AB	Abstract
PT	Publication type
EM	Email addresses

These columns were chosen on the basis of their importance in terms of semantic content as well as the fact that they tended to be the most populated of all the WOK/WOS fields. Of special note are the AU, EM, and AB fields.

Authors and Email Addresses Dimensions

Within the WOK/WOS schema (as mentioned previously), AU and EM information is encoded into a semicolon separated list. Thus given two authors for a paper A1, A2, the AU field would contain: "A1; A2." In order to properly analyze this information, it must be decoded and each separate entity is inserted individually into the dimension table.

Abstract Dimension

In order to process the occurrence of certain words within the abstract, the entire abstract of each WOK record is treated as a collection of individual terms, separated by blank space, punctuation, and parentheses. Using the same concept as with the AU and EM dimensions, each word and it's associated WOK record are stored in the dimension table.

All together, the data warehouse for the example implementation is depicted in Fig. 3. The number of rows in the fact table (which depends on the granularity of the dimensional tables) can grow to be quite large. For instance, a WOS search that returned 296 rows, which, when properly dimensioned, produced a fact table consisting of 253,067 rows.

5 Conclusion

BI tools and techniques can greatly simplify the analysis of large amounts of data. By utilizing a common schema format defined in XML, the "WOSToDB" tool is able to serialize any conforming data to an SQL database. From this, data is constructed an online analytical processing cube, which can then be used to quickly and efficiently analyze the data. While the implementation of a BI project requires a certain amount of knowledge about data management tools and techniques, the end result is a reusable system able to provide efficient and complex data analysis suitable to many varied problem domains.

References

1. Business intelligence: Wikipedia, the free encyclopedia. Page Version ID: 540682028 (February 2013)
2. Dimension (data warehouse): Wikipedia, the free encyclopedia. Page Version ID: 536574434 (February 2013)
3. Rose, M.T., Hollenbeck, S., Masinter, L.: Guidelines for the use of extensible markup language (XML) within IETF protocol. http://tools.ietf.org/html/rfc3470 (February 2013)
4. XML: Wikipedia, the free encyclopedia. Page Version ID: 539777964 (February 2013)
5. XSLT: Wikipedia, the free encyclopedia. Page Version ID: 540812551 (February 2013)

New Exceptional Orthogonal Polynomials (EOPs) and Nonlinear Algebras Associated to the Quantum System

D. Dutta

Abstract Past few years have witnessed a considerable level of research activity in the field of exceptional orthogonal polynomials (EOPs), which are new complete orthogonal polynomial systems, and these are first observed as a result of the development of a direct approach to exact or quasi-exact solvability for spectral problems in quantum mechanics that would go beyond the classical Lie algebraic formulations. We have discovered new EOP families associated to such kind of above systems in the framework of supersymmetric quantum mechanics. We have studied thoroughly some fundamental properties of those EOP families. We also have been able to prove completeness of few such EOP categories in weighted Hilbert space, associated with solutions of certain conditionally exactly solvable potentials obtained via unbroken as well as broken supersymmetry. Some important key properties of such polynomials, e.g, recurrence relation, Rodrigues formula, ladder operators, differential equations, etc., have been obtained.

1 Introduction

It is known to us that the classical orthogonal polynomials (COP), i.e., Hermite, Jacobi, Laguerre, etc., have wide application in applied mathematics and physics. Especially bound state solutions of some standard quantum mechanical problems admit COPs. On the other hand, one of the most interesting development in recent years is to construct new exactly solvable potentials in connection with the appearance of families of exceptional orthogonal polynomials (EOP). We aimed at re-examining some earlier results [1, 2] related to conditionally exactly solvable potentials [3, 4] in the regime of EOPs, supersymmetry and polynomial algebras. In our study, we have considered solutions of conditionally exactly solvable partners of the radial and the linear oscillator potential with broken as well as unbroken supersymmetry. More interestingly, the polynomial algebras have been treated over the whole/part

D. Dutta (✉)
Physics & Applied Mathematics Unit, Indian Statistical Institute,
Kolkata 700 108, India
e-mail: debjitmath@yahoo.co.in

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_29

of the space of such EOPs. Some important key properties of these polynomials, e.g, generating function, Rodrigues type formula, etc., could have found. Moreover, we applied new approach to demonstrate completeness of few EOP families in the associated Hilbert space.

2 EOPs Associated to Broken Supersymmetry

We recall that a pair of Hamiltonians for spherical oscillator of the form [5, 6]

$$\begin{aligned} H_\pm &= A^\pm A^\mp = \frac{1}{2}\left[-\frac{d^2}{dr^2} + V_\pm(r)\right] \\ V_\pm(r) &= W^2(r) \pm W'(r) \end{aligned} \tag{1}$$

where $A^\pm$ are described by

$$A^+ = \frac{1}{\sqrt{2}}\left(\frac{d}{dr} + W(r)\right), \quad A^- = \frac{1}{\sqrt{2}}\left(-\frac{d}{dr} + W(r)\right) \tag{2}$$

form a supersymmetric system and by construction the above Hamiltonians are isospectral except perhaps the zero energy state (which if it exists is assumed to belong to H_-). In this case, supersymmetry is referred to be unbroken and the relationship between the energies and the eigenfunctions of these Hamiltonians are given by

$$E_0^- = 0, \quad E_{n+1}^- = E_n^+ > 0 \tag{3}$$

$$\psi_0^- = N\, e^{-\int W(r)dr}, \quad \psi_n^+ = \frac{1}{\sqrt{E_{n+1}^-}}\, A^+\psi_{n+1}^-, \quad \psi_{n+1}^- = \frac{1}{\sqrt{E_n^+}}\, A^-\psi_n^+. \tag{4}$$

In other words the zero energy ground state is a singlet while the excited states are doubly degenerate. On the other hand, if neither of $\psi_0^\pm = e^{\pm\int W(r)dr}$ is normalizable, then supersymmetry is broken and we have

$$E_n^+ = E_n^- > 0, \quad \psi_n^+ = \frac{1}{\sqrt{E_n^-}}\, A^+\psi_n^-, \quad \psi_n^- = \frac{1}{\sqrt{E_n^+}}\, A^-\psi_n^+. \tag{5}$$

Thus, in this case the ground as well as the excited states are doubly degenerate.

Here, we shall consider isospectral partner of the radial oscillator system. Let us first consider the case of broken supersymmetry. In this case, the superpotential is given by [2]

$$W(r) = r + \frac{\gamma+1}{r} + \frac{u'}{u}, \quad 0 < r < \infty \tag{6}$$

where $u(r^2)$ is suggested as

$$u(r^2) = {}_1F_1\left(\frac{1-\epsilon}{2}, \gamma + \frac{3}{2}, -r^2\right). \tag{7}$$

It can be easily verified that neither of $\psi_0^{\pm} = (u)^{\pm 1} r^{\pm(\gamma+1)} e^{\pm\frac{r^2}{2}}$ is normalizable which lead to the fact that supersymmetry is spontaneously broken. In this case the partner potentials are given by

$$V_+(r) = \frac{r^2}{2} + \frac{\gamma(\gamma+1)}{2r^2} + \epsilon + \gamma + \frac{1}{2} \tag{8}$$

$$V_-(r) = \frac{r^2}{2} + \frac{(\gamma+1)(\gamma+2)}{2r^2} + \frac{u'(r^2)}{u(r^2)}\left(2r + 2\frac{\gamma+1}{r} + \frac{u'(r^2)}{u(r^2)}\right) - \epsilon + \gamma + \frac{3}{2}. \tag{9}$$

The potential in (8) represents the standard radial oscillator potential whose energy and eigenfunctions are given by

$$E_n^+ = 2n + 2\gamma + 2 + \epsilon, \quad \psi_n^+ = \sqrt{\frac{2(n)}{\Gamma(n+\gamma+\frac{3}{2})}}\, r^{\gamma+1} L_n^{\gamma+\frac{1}{2}}(r^2)\, e^{-\frac{r^2}{2}}, \tag{10}$$
$$n = 0, 1, 2, \ldots .$$

Note that in this case V_- is a nonshape-invariant potential (or more precisely a conditionally exactly solvable one [3, 4]) and has the same spectrum as V_+. It's eigenfunctions may be obtained using (5) and are given by

$$\psi_n^-(r) = \sqrt{\frac{2(n)}{(4n+4\gamma+4+2\epsilon)\Gamma(n+\gamma+\frac{3}{2})}}\, \frac{e^{-\frac{r^2}{2}} r^{\gamma+2}}{u(r^2)} \left[\frac{u'(r^2)}{r} L_n^{\gamma+\frac{1}{2}}(r^2) + 2u(r^2) L_n^{\gamma+\frac{3}{2}}(r^2)\right]. \tag{11}$$

Now we intend to identify the expression inside the square bracket as the EOP $p_n(r^2)$ and the prefactor as the square root of the weight function $w(r^2)$, i.e,

$$p_n(r^2) = \left[\frac{u'(r^2)}{r} L_n^{\gamma+\frac{1}{2}}(r^2) + 2u(r^2) L_n^{\gamma+\frac{3}{2}}(r^2)\right] \tag{12}$$

$$w(r^2) = \frac{e^{-r^2} r^{2\gamma+4}}{u^2(r^2)}. \tag{13}$$

It is clear that for $p_n(r^2)$ to be a polynomial, one has to choose the parameter ϵ such that $u(r^2)$ is a polynomial. First few members of this family are given by

$$\begin{aligned} p_0(r^2) &= \frac{u'(r^2)}{r} + 2u(r^2), \\ p_1(r^2) &= \frac{1}{2}[(2\gamma + 3 - 2r^2)\frac{u'(r^2)}{r} + 2(2\gamma + 5 - 2r^2)u(r^2)] \end{aligned} \tag{14}$$

We shall now explore some properties of these polynomials which can be studied without having to specify γ and ϵ. First we note that

$$\langle \psi_m^+ | \psi_n^+ \rangle = \delta_{mn} \tag{15}$$

and consequently using (5) we find that

$$\int_0^\infty w(r^2) p_m(r^2) p_n(r^2) dr = \frac{(2n + 2\gamma + 2 + \epsilon)\Gamma(n + \gamma + \frac{3}{2})}{n} \delta_{mn}, \tag{16}$$

i.e, the polynomials $p_n(r^2)$ are orthogonal with respect to weight function $w(r^2)$ on the positive half line. We obtain a closed form of the generating function for the EOPs (12)

$$F(r, z) = \frac{e^{\frac{r^2 z}{z-1}}}{(1 - z)^{\gamma+5/2}} \left[(1 - z)\frac{u'(r^2)}{r} + 2u(r^2) \right] \tag{17}$$

Now we shall obtain another result, namely, a Rodrigues type formula for these polynomials. Before obtaining this it may be noted that there exists quantum number independent raising ($L^\dagger$) and lowering operators (L) for the wave functions $\psi_n^-(r)$ given by [2, 7]

$$L = A^\dagger c A, \quad L^\dagger = A^\dagger c^\dagger A, \tag{18}$$

where $c = a^2 - \frac{\gamma(\gamma+1)}{2r^2}$, $c^\dagger = (a^+)^2 - \frac{\gamma(\gamma+1)}{2r^2}$, and $a^\dagger, a$ being the standard harmonic oscillator raising and lowering operators while $A, A^\dagger$ are defined by (2).

Now denoting the raising and lowering operators for $p_n(r)$ by $\mathcal{L}^\dagger$ and $\mathcal{L}$ it can be shown using (5) that

$$\begin{aligned} \mathcal{L} p_n(r^2) &= -2(2n + 2\gamma + 2 + \epsilon)\sqrt{(2n + 2\gamma + 1)(2n + 2\gamma + 3)}\; p_{n-1}(r^2) \\ \mathcal{L}^\dagger p_n(r^2) &= -2(n + 1)(2n + 2\gamma + 2 + \epsilon)\sqrt{\tfrac{2n+2\gamma+5}{2n+2\gamma+3}}\; p_{n+1}(r^2) \end{aligned} \tag{19}$$

It can be seen that the ladder operators do not depend on the order of the polynomials. From (19) it follows that

$$p_n(r^2) = \left(-\frac{1}{4}\right)^n \frac{\Gamma(\gamma + 1 + \frac{\epsilon}{2})}{n\Gamma(n + \gamma + 1 + \frac{\epsilon}{2})} \sqrt{\frac{(\gamma + \frac{3}{2})}{(n + \gamma + \frac{3}{2})}}\, (\mathcal{L}^\dagger)^n p_0(r^2). \tag{20}$$

The relation (20) is a Rodrigues type formula for the EOPs $p_n(r^2)$.

It is well known that different Lie algebras can be realized in the space of orthogonal polynomials. Here, it will be shown that a type of cubic algebra can be realized over space of EOPs. It can be shown that the operators $\mathcal{L}^\dagger$, $\mathcal{L}$, and h satisfy the following commutation relations:

$$\begin{aligned}
[\mathcal{L}^\dagger, h] &= 4\mathcal{L}^\dagger \\
[\mathcal{L}, h] &= -4\mathcal{L} \\
[\mathcal{L}, \mathcal{L}^\dagger] &= -h[2(h+4\gamma+2\epsilon+4)^2 - (h+4\gamma+2\epsilon+4)(2\epsilon+10\gamma+9) \\
&\quad +4\gamma\epsilon+10\epsilon+8\gamma^2+36\gamma+40].
\end{aligned} \tag{21}$$

2.1 $\epsilon = 3$

Consider simplest possible $\epsilon = 3$, the lowest value of ϵ for which u given by (7) is a polynomial. In this case using (8) we find that

$$V_+(r) = \frac{r^2}{2} + \frac{\gamma(\gamma+1)}{r^2} + \gamma + \frac{7}{2} \tag{22}$$

$$V_-(r) = \frac{r^2}{2} + \frac{(\gamma+1)(\gamma+2)}{2r^2} - \frac{4}{2r^2+2\gamma+3} + \frac{16r^2}{(2r^2+2\gamma+3)^2} + (2\gamma+5). \tag{23}$$

Also from (7) we get[1]

$$u(x) = \frac{1}{(\gamma+\frac{3}{2})}\left(x+\gamma+\frac{3}{2}\right). \tag{24}$$

From (12) some members of the polynomial family can be found to be

$$\begin{aligned}
p_0(x) &= \frac{1}{(\gamma+\frac{3}{2})}(2x+2\gamma+5), \\
p_1(x) &= \frac{1}{(\gamma+\frac{3}{2})}\left(2\gamma^2+10\gamma+\frac{21}{2}-2x^2\right).
\end{aligned} \tag{25}$$

Also the weight function in this case becomes

$$w(x) = \frac{1}{2}\frac{e^{-x}x^{\gamma+\frac{3}{2}}}{u^2(x)} \tag{26}$$

[1] We now consider EOPs in terms of the variable $x = r^2 \in [0, \infty)$.

3 Conclusion

We have been able to invent new EOPs and study various properties of two such types of EOP families associated to some conditionally exactly solvable but nonshape-invariant system, namely supersymmetric partner potentials of the radial and linear oscillator potentials. For specific choices of ϵ we would get the EOP. Interestingly, EOPs may be reproduced with the help of higher-order Darboux transformation [8, 9, 10]. Point to be noted that, in general, ladder operators for the EOPs may not be obtained unless the symmetry of the original problem (in the present case it is the radial oscillator potential) is known. Once if it is known then one can establish various formulas, e.g, Rodrigues type formula without much formidability. In such cases the polynomials form the representation space of polynomial algebras. We would like to point out that in the present case the ladder operators are independent of the order of polynomial. We have observed that the ladder operators and the differential operator for the polynomial may still form a cubic algebra, emerging from a particular physical system, namely the radial oscillator.

References

1. Dubov, S. Y., Eleonskii, V. M., Kulagin, N.E.: Equidistant spectra of anharmonic oscillators. Sov. Phys. JETP. 75(3):446–451 (1992); Dubov, S. Y., Eleonskii, V. M., Kulagin, N.E.: Equidistant spectra of anharmonic oscillators. Chaos 4(47):67–72 (1994)
2. Junker, G., Roy, P.: Conditionally exactly solvable potentials: a supersymmetric construction method. Ann. Phys. 270(1):155–177 (1998)
3. de Souza Dutra, A.: Conditionally exactly soluble class of quantum potentials. Phys. Rev. A 47(4):R2435–R2437 (1993)
4. Dutt, R., Khare, A., Varshni, Y.P.: New class of conditionally exactly solvable potentials in quantum mechanics. J. Phys. A 28(11):L107–L113 (1995)
5. Cooper, F., Khare, A., Sukhatme, U.: Supersymmetry in quantum mechanics. World Scientific, Singapore, (2001)
6. Junker, G.: Supersymmetric methods in quantum and statistical physics. Springer-Verlag, London, (1996)
7. Dutta, D., Roy, P.: Conditionall exactly solvable potentials and exceptional orthogonal polynomials. J. Math. Phys. 51(4):042101–042110 (2010)
8. Samsonov, B.F., Ovcharov, I.N.: Darboux transformation for the nonsteady Schrödinger equation. Russ. Phys. J. 38(7):706–712 (1995)
9. Adler, V.E.: A modification of Crum's method. Theor. Math. Phys. 101(3):1381–1386 (1994)
10. Bagrov, V.G., Samsonov, B.F.: Darboux transformation and elementary exact solutions of the Schrödinger equation. Pramana. 49(6):563–571 (1997)

Avoiding the Coordinate Singularity Problem in the Numerical Solution of the Dirac Equation in Cylindrical Coordinates

F. Fillion-Gourdeau, E. Lorin and A.D. Bandrauk

Abstract A new numerical method is developed for the solution of the Dirac equation for 3D axisymmetric geometries using cylindrical coordinates. It is based on a split-step scheme in coordinate space, which can be parallelized very efficiently. A new technique to circumvent the coordinate singularity at $r = 0$ using Poisson's integral solution of the wave equation for the radial operator is used. The general strategy is to interpolate the solution using cubic Hermite polynomials and to integrate exactly the Poisson solution. The result of this procedure gives a nonstandard finite difference scheme on a time staggered grid. The numerical method is then utilized to evaluate the ground state of an electron bound in a Coulomb potential.

1 Introduction

The Dirac equation is one of the pillars of theoretical physics as it describes relativistic fermions, which are ubiquitous in nature. As a consequence, this equation is required in the theoretical investigation of many observables in quantum mechanics and quantum field theory such as electron–positron production, vacuum polarization, heavy molecules spectra, molecular ionization rates, and many others. Its wide range of applicability is now well established and it now finds applications in many areas of physics. However, it is well known that solving this equation is a challenging

F. Fillion-Gourdeau (✉)
Centre de Recherches Mathématiques, Université de Montréal,
Montréal H3T 1J4, Canada
e-mail: filliong@CRM.UMontreal.ca

E. Lorin
School of Mathematics and Statistics, Carleton University,
Ottawa K1S 5B6, ON, Canada
e-mail: elorin@math.carleton.ca

A. D. Bandrauk
Laboratoire de chimie théorique, Faculté des Sciences,
Université de Sherbrooke, Sherbrooke J1K 2R1, QC, Canada
e-mail: andre.bandrauk@usherbrooke.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_30

problem, both from the analytical and the numerical sides. Therefore, a lot of efforts were devoted to the development of new numerical methods to solve this equation. For instance, existing numerical schemes have been studied in [1–8] and include different approaches such as split-operator, spectral, finite element, and finite difference methods.

Recently, a new split-step scheme was developed based on the method of characteristics [9, 10]. It was demonstrated that this scheme can be parallelized very efficiently. The main goal of this chapter is to present an extension of this numerical scheme to cylindrical coordinates and new applications for physical systems having an azimuthal symmetry.

2 Numerical Method

The main equation considered in this work is the Dirac equation in cylindrical coordinates, which is given by

$$\begin{aligned} i\partial_t \psi(t,r,z) =& \Big\{ \alpha_x \Big[-ic\partial_r - ic\frac{1}{2r} - eA_r(t,r,z) \Big] + \alpha_y \Big[c\frac{j_z}{r} - eA_\theta(t,r,z) \Big] \\ &+ \alpha_z \Big[-ic\partial_z - eA_z(t,r,z) \Big] + \beta mc^2 + eV(t,r,z) \Big\} \psi(t,r,z). \qquad (1) \end{aligned}$$

where the radial distance is $r = \sqrt{x^2+y^2} \in \mathbb{R}^+$. Also, we have $\psi(t,r,z) \in L^2(\mathbb{R}^+,\mathbb{R}^+,\mathbb{R}) \otimes \mathbb{C}^4$ is the time and coordinate-dependent four-spinor, $\mathbf{A}(t,r,z)$ represents the three space components of the electromagnetic vector potential, $V(t,r,z) = A_0(t,r,z)$ is the scalar potential, e is the electric charge, $\mathbb{I}_n$ is the n by n unit matrix and α_i, β are the Dirac matrices. Finally, $j_z = ..., -\frac{5}{2}, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, ...$ is the angular momentum, obtained by factorizing the azimuthal part (we are assuming here that the potential does not depend on θ, the polar angle). This equation describes physically the relativistic dynamics of a single electron subject to an external electromagnetic field with azimuthal symmetry. The Dirac matrices are

$$\alpha_i := \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \ , \ \beta := \begin{bmatrix} \mathbb{I}_2 & 0 \\ 0 & -\mathbb{I}_2 \end{bmatrix}. \qquad (2)$$

The σ_i are the usual 2 × 2 Pauli matrices.

We are then interested in the Cauchy problem where Eq. (1) is solved with the initial condition $\psi(t_n,r,z) = \psi^n(r,z)$. An operator splitting scheme is used to reach this goal, which is given by

$$\begin{aligned} i\partial_t \psi^{(1)}(t) &= \hat{A}\psi^{(1)}(t), \quad \psi^{(1)}(t_n) = \psi^n, \quad && t \in [t_n, t_{n+1}) \\ i\partial_t \psi^{(2)}(t) &= \hat{B}\psi^{(2)}(t), \quad \psi^{(2)}(t_n) = \psi^{(1)}(t_{n+1}), \quad && t \in [t_n, t_{n+1}) \\ i\partial_t \psi^{(3)}(t) &= \hat{D}\psi^{(3)}(t), \quad \psi^{(3)}(t_n) = \psi^{(2)}(t_{n+1}), \quad && t \in [t_n, t_{n+1}) \\ \text{and} \quad \psi^{n+1} &= \psi^{(3)}(t_{n+1}) \end{aligned} \qquad (3)$$

where the upper subscript in parenthesis on the wave function denotes the splitting step number. The numerical error scales like $O(\delta t^2)$ where we denote $\delta t := t_{n+1} - t_n$ (higher order can also be obtained [10]). The operators are defined as

$$\hat{A} := -ic\alpha_x \partial_r - ic\alpha_x \frac{1}{2r} + c\alpha_y \frac{j_z}{r}, \tag{4}$$

$$\hat{B} := -ic\alpha_z \partial_z, \tag{5}$$

$$\hat{D} := \beta mc^2 + e\mathbb{I}_4 V(t,r,z) - e\alpha_x A_r(t,r,z) - e\alpha_y A_\theta(t,r,z) - e\alpha_z A_z(t,r,z). \tag{6}$$

The second step of the splitting can be solved exactly using the method of characteristics [9, 10]. Then, by choosing a space discretization such that $c\delta t = \delta z$, the time evolution for this step is exact and given by

$$\psi^{(2)}(t_{n+1}, j, k) = \frac{1}{2}\left\{[\mathbb{I}_4 + \alpha_z]\psi^{(2)}(t_n, j, k-1) + [\mathbb{I}_4 - \alpha_z]\psi^{(2)}(t_n, j, k+1)\right\}. \tag{7}$$

Here, j, k label the points on the grid for r and z coordinates, respectively.

In the last step of the splitting, there also exists a formal solution given in terms of a time-ordered exponential. The latter can be approximated to an order of accuracy $O(\delta t^3)$ by

$$\psi^{(3)}(t_{n+1}, j, k) = U(j,k)\exp\left[-ieV^{n+\frac{1}{2}}(j,k)\right]\psi^{(3)}(t_n, j, k), \tag{8}$$

where

$$U(j,k) = e^{-ie\delta t\left[\beta\frac{mc^2}{e} - \alpha_x A_r^{n+\frac{1}{2}}(j,k) - \alpha_y A_\theta^{n+\frac{1}{2}}(j,k) - \alpha_z A_z^{n+\frac{1}{2}}(j,k)\right]} \tag{9}$$

is a 4 × 4 matrix that can be evaluated explicitly.

There also exists an analytical solution for the first step of the splitting. First, the spinor components ($a = 1, \cdots, 4$, $b = 4, \cdots, 1$) have to be decoupled, which yields the following Cauchy problem

$$\begin{cases} \partial_t^2 \psi_a^{(1)}(t,r,z) = c^2\left[\partial_r^2 + \frac{1}{r}\partial_r - \frac{\mu_a^2}{r^2}\right]\psi_a^{(1)}(t,r,z), \\ \psi_a^{(1)}(t_n,r,z) = g_a(r,z), \quad \partial_t\psi_a^{(1)}(t_n,r,z) = h_a(r,z) = c\left[-\partial_r - \frac{\mu_b}{r}\right]g_b(r,z) \end{cases} \tag{10}$$

where $\mu_a = \mu_{1,2}$ for $a = 1,3$ and $a = 2,4$, respectively, and $\mu_b = -\mu_1$ for $b = 1,3$ and $\mu_b = \mu_2$ for $b = 2,4$. One immediately recognizes the wave equation in polar coordinates which has an integral solution given by Poisson's formula [11]

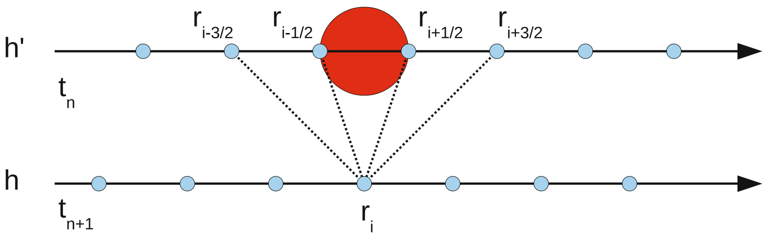


Fig. 1 Description of the staggered mesh in radial coordinates. The *red circle* is the integration region B. The *lines* represent the stencil of the scheme

$$\psi_a^{(1)}(t_{n+1}, r, z) = \frac{1}{2\pi c\delta t}\int_{B(r,c\delta t)} RdRd\theta \frac{1}{\sqrt{c^2\delta t^2 - [R^2 + r^2 - 2Rr\cos(\theta)]}} \\ \times \Big\{\cos(\mu_a\theta)\,[g_a(R,z) + \delta t h_a(R,z) + [R - r\cos(\theta)]\partial_R g_a(R,z)] \\ - \sin(\mu_a\theta)\frac{r}{R}\mu_a\sin(\theta)g_a(R,z)\Big\}. \tag{11}$$

Here, the integration region $B(r, c\delta t)$ is a disk of radius $c\delta t$ centered at r in the $r - \theta$-plane (it is depicted in Fig. 1). The last part of this section is devoted to the approximation of this integral. The strategy to perform this task is now summarized:

- At $t = t_n$, the grid points are chosen on the boundaries of the integration region B, in the radial direction, with $\delta r = 2c\delta t$. Automatically, the grid points at $t = t_{n+1}$ will be staggered (see Fig. 1) to be consistent with Poisson's formula, which yields the value of the wave function at the center of the integration region.
- An approximation of the wave function at $t = t_n$ between grid points is obtained by cubic Hermite polynomial interpolation. This is required to perform the integral on the integration region.
- Substituting the polynomial approximation in Poisson's formula, we obtain integrals of this form:

$$I_l(r) = \int_{-\theta_{\max}}^{\theta_{\max}} d\theta \int_{R^-}^{R^+} dR \frac{f(\theta)R^l}{\sqrt{a^2 - [R^2 + r^2 - 2Rr\cos(\theta)]}}, \tag{12}$$

where l is an integer while $\theta_{\max}$ and $R^{\pm}$ characterize the integration region B. These integrals can be computed analytically and were implemented on a symbolic algebra language software.

The resulting numerical scheme can then be written in a form reminiscent of a nonstandard finite difference scheme (this is depicted in Fig. 1):

$$\begin{aligned}\psi_a^{(1)}(t_{n+1}, j, k) =& A_1(j)\psi_a^{(1)}(t_n, j-3/2, k) + A_2(j)\psi_a^{(1)}(t_n, j-1/2, k)\\ &+ A_3(j)\psi_a^{(1)}(t_n, j+1/2, k) + A_4(j)\psi_a^{(1)}(t_n, j+3/2, k)\\ &+ B_1(j)\psi_b^{(1)}(t_n, j-3/2, k) + B_2(j)\psi_b^{(1)}(t_n, j-1/2, k)\\ &+ B_3(j)\psi_b^{(1)}(t_n, j+1/2, k) + B_4(j)\psi_b^{(1)}(t_n, j+3/2, k) \qquad (13)\end{aligned}$$

where the finite difference coefficients A, B can be evaluated explicitly and depend on the radial position. Also, this scheme is well defined at $r = 0$, even if the Dirac operator had singular terms (like $1/r$).

3 Numerical Results

The first test being considered in this study concern Gaussian wave packets, where the initial wave function is given by $\psi_1(t = 0, r) = r^{|\mu_1|} e^{-\frac{r^2}{4\Delta^2}}$, where Δ characterizes the Gaussian width. Physically, this corresponds to a free electron. The results are shown in Fig. 2 for the time evolution of the wave packet (with a comparison to an analytical solution) and the order of convergence. The next benchmark tests concern the time evolution of bound states of the regularized Coulomb potential[1]. It is possible to compute time-independent wave functions from a time-dependent numerical scheme by using the well-known Feit–Fleck method [12]. The bound state of a hydrogen-like atom is given in Fig. 2 along with its power spectrum obtained from the Feit–Fleck method. The ground state energy is $E_{\text{ground}} \approx 18710.3$ a.u.. This value is close to the analytical Coulomb ground state energy given by $E_{\text{ground}} \approx 18729.9$, with a relative difference of $\delta_{\text{rel}} \approx 0.1\,\%$. This numerical ground state can now serve as an initial state for the study of relativistic ionization by adding an external laser field. This will be the topic of future investigations.

[1] The Coulomb singularity is regularized by using a constant distribution of charge inside the nucleus

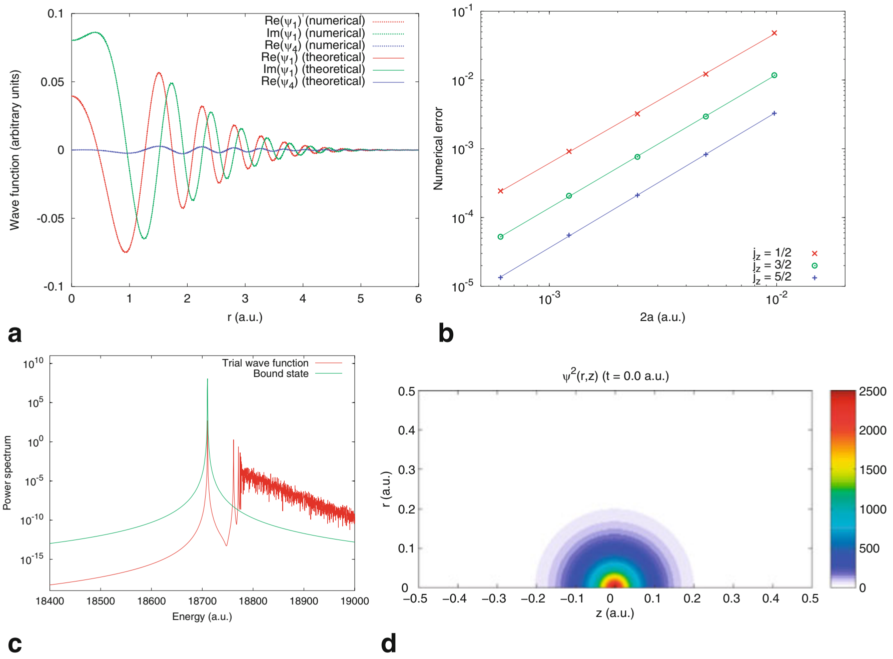


Fig. 2 **a** Results for the nonzero components of the Gaussian wave packet with an initial width of $\Delta = 0.1$, evaluated at time $t = 0.143$ a.u. The theoretical and calculated curves overlap. **b** The order of convergence is 1.8999 for $j_z = 1/2$, 1.9435 for $j_z = 3/2$ and 1.9767 for $j_z = 5/2$. **c** Results for the power spectrum of trial wave function and the bound state constructed from the Feit–Fleck method. **d** Ground state of a hydrogen-like atom with $Z = 10$

References

1. Braun, J.W., Su, Q., Grobe, R.: Numerical approach to solve the time-dependent Dirac equation. Phys. Rev. A **59**(1), 604–612 (1999)
2. Mocken, G.R., Keitel, C.H.: Quantum dynamics of relativistic electrons. J. Comput. Phys. **199**(2), 558–588 (2004)
3. Momberger K., Belkacem, A., Sørensen, A.H.: Numerical treatment of the time-dependent Dirac equation in momentum space for atomic processes in relativistic heavy-ion collisions. Phys. Rev. A **53**(3), 1605–1622 (1996)
4. Huang, Z., Jin, S., Markowich, P.A., Sparber, C., Zheng, C.: A time-splitting spectral scheme for the Maxwell–Dirac system. J. Comput. Phys. **208**(2), 761–789 (2005)
5. Bao, W., Li, X.G.: An efficient and stable numerical method for the Maxwell–Dirac system. J. Comput. Phys. **199**(2), 663–687 (2004)
6. Bottcher, C., Strayer, M.R.: Numerical solution of the time-dependent Dirac equation with application to positron production in heavy-ion collisions. Phys. Rev. Lett. **54**(7), 669–672 (1985)
7. Becker, U., Grun, N., Scheid, W.: Solution of the time-dependent Dirac equation by the finite difference method and application for $Ca^{20+} + U^{91+}$. J. Phys. B: At. Mol. Phys. **16**(11), 1967 (1983)

8. Selstø, S., Lindroth, E., Bengtsson, J.: Solution of the Dirac equation for hydrogen-like systems exposed to intense electromagnetic pulses. Phys. Rev. A **79**(4), 043418 (2009)
9. Lorin, E., Bandrauk, A.D.: A simple and accurate mixed P0-Q1 solver for the Maxwell-Dirac equations. Nonlinear Anal. Real World Appl. **12**(1), 190–202 (2011)
10. Fillion-Gourdeau, F., Lorin, E., Bandrauk, A.D.: Numerical solution of the time-dependent dirac equation in coordinate space without fermion-doubling. Comput. Phys. Commun. **183**(7), 1403–1415 (2012)
11. Evans, L.C.: Partial differential equations. Graduate studies in mathematics. Am. Math. Soc., Providence (1997)
12. Feit, M.D., Fleck, J.A., Steiger, A.: Solution of the Schrodinger equation by a spectral method. J. Comput. Phys. **47**(3), 412–433 (1982)

Symmetry Reductions and Exact Solutions of a Generalized Fisher Equation

M. L. Gandarias, M. Rosa and M. S. Bruzon

Abstract In this chapter, we study a generalized Fisher equation based on the theory of symmetry reductions in partial differential equations. Optimal systems and reduced equations are obtained. We derive some travelling wave solutions by applying the (G'/G)-expansion method to one of these reduced equation.

1 Introduction

The Fisher–Kolmogorov equation, proposed for population dynamics in 1930, shows the spread of an advantageous gene into a population. As described by Britton [2], generalizations of this equation are needed to more accurately model complex diffusion and reaction effects found in many biological systems. The equation analyzed in this chapter is a generalized Fisher equation in which $g(u)$ is the diffusion coefficient depending on the variable u, x and t being the independent variables, and $f(u)$ an arbitrary function

$$u_t = f(u) + (g(u)u_x)_x \tag{1}$$

Equation (1) is also known as the density-dependent diffusion-reaction equation which is mentioned by J. D. Murray in [7]. Reaction-diffusion equations arise from modeling densities of particles such as substances and organisms which disperse through space as a result of the irregular movement of every particle.

Due to the interest of these equations, a lot of attention has been paid to the use of Lie point symmetry methods to exploit the invariance of the generalized equation [4] and references therein. In [5], we determined the subclasses of these equations which are nonlinear self-adjoint. By using a general theorem on conservation laws proved

M. L. Gandarias (✉) · M. Rosa · M. S. Bruzon
University of Cadiz Puerto Real, Puerto Real, Spain
e-mail: marialuz.gandarias@uca.es

M. Rosa
e-mail: maria.rosa@uca.es

M. S. Bruzon
e-mail: m.bruzon@uca.es

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_31

by Nail Ibragimov and the symmetry generators we found conservation laws for these partial differential equations. There is no existing general theory for solving nonlinear partial differential equations (PDEs) and the machinery of the Lie group theory provides the systematic method to search for the special group-invariant solutions. The knowledge of the optimal system of subalgebras gives the possibility to construct the optimal system of solutions and permits the generation of new solutions starting from invariant or noninvariant solutions.

Due to the great advance in computation in the last few years, a great progress has been made in the development of methods and their applications for finding solitary travelling-wave solutions of nonlinear evolution equations [3, 9, 6]. For (1) the list of nontrivial Lie generators were derived in [4] by combining the standard method of group classification and the form-preserving transformation.

The aim of this chapter is to study the density-dependent diffusion-reaction Eq. (1) from the point of view of the theory of symmetry reductions in partial differential equations. We construct the reductions from the optimal system of subalgebras.

Then, due to the fact that Eq. (1) admits groups of space and time translations, we search for travelling wave solutions of the density-dependent diffusion-reaction Eq. (1), with physical interest, when the diffusion coefficient $g(u)$ follows a power law. In order to do that we apply the well known $\frac{G'}{G}$-expansion method [3, 9], to the reduced equation.

2 Symmetry Reductions

The Lie classical method applied to (1) yields (see [4]):
For $f(u)$ and $g(u)$ arbitrary, *the only symmetry that is admitted* by (1) is

$$\mathbf{v}_2 = \frac{\partial}{\partial t}.$$

For some special choices of the functions $f(u)$ and $g(u)$ it can also be extended to the cases listed below:

1. For $f(u) = u^m$ and $g(u) = u^n (m \neq n + 1)$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_1 = x\frac{\partial}{\partial x} + \frac{2(m-1)}{n-m+1}\frac{\partial}{\partial t} + \frac{2u}{n-m+1}\frac{\partial}{\partial u}$$

2. For $f(u) = \frac{u^{n+1}}{n+1}$ and $g(u) = u^n$ ($n \neq 0$ and $n \neq -1$) we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_3 = nt\frac{\partial}{\partial t} - u\frac{\partial}{\partial u}$$

3. For $f(u) = \frac{3c_1 u}{4} + \frac{c_2}{u^{\frac{1}{3}}}$ and $g(u) = u^{-\frac{4}{3}}$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_4 = 4\,e^{c_1 t}\frac{\partial}{\partial t} + 3\,c_1 e^{c_1 t} u\frac{\partial}{\partial u}$$

4. For $f(u) = -\frac{c_1 u}{n}$ and $g(u) = u^n$ $(n \neq 0)$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_5 = nx\frac{\partial}{\partial x} + 2u\frac{\partial}{\partial u}, \quad \mathbf{v}_6 = n\, e^{c_I t}\frac{\partial}{\partial t} - c_I e^{c_I t} u\frac{\partial}{\partial u}$$

5. For $f(u) = c_1 u$ and $g(u) = u^{-1}$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_5, \mathbf{v}_6, \mathbf{v}_7 = (x \log(x) - x)\,\tfrac{\partial}{\partial x} - 2u\log(x)\tfrac{\partial}{\partial u}$$

6. For $f(u) = c_2 e^{nu} - \frac{c_I}{n}$ and $g(u) = de^{nu}$ $(n \neq 0\ )$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_8 = ne^{c_I t}\frac{\partial}{\partial t} - c_I e^{c_I t}\frac{\partial}{\partial u}$$

7. For $f(u) = -\frac{c_I}{n}$ and $g(u) = de^{nu}$ $(n \neq 0)$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_8, \mathbf{v}_9 = nx\frac{\partial}{\partial x} + 2\frac{\partial}{\partial u}$$

8. For $f(u) = c_2 e^{nu}$ and $g(u) = de^{nu}(n \neq 0)$ we obtain the following generators:

$$\mathbf{v}_2, \mathbf{v}_{10} = nt\frac{\partial}{\partial t} - \frac{\partial}{\partial u}$$

2.1 Optimal Systems and Reductions

The corresponding generators of the optimal system of subalgebras, [8] are:

1. For $f(u) = u^n$ and $g(u) = u^m$

$$a\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_3.$$

2. For $f(u) = e^{nu}$ and $g(u) = e^{mu}$

$$a\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_4.$$

3. For $f(u) = c_2 u^{n+1} - \frac{c_1 u}{n}$ and $g(u) = u^n$

$$a\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_5.$$

4. For $f(u) = u^{-\frac{1}{3}}$ and $g(u) = u^{-\frac{4}{3}}$

$$a\mathbf{v}_1 + \mathbf{v}_2, b\mathbf{v}_1 + \mathbf{v}_3, c\mathbf{v}_3 + \mathbf{v}_6, d\mathbf{v}_3 + \mathbf{v}_7,$$

where $a, b, c, d \in R$ are arbitrary.

In the following, reductions of Eq. (1) to ordinary differential equations (ODEs) are obtained by using the generators of the optimal system.

Reduction 1 Generator $\mathbf{v}_1 + \mathbf{v}_2$. Substituting the similarity variable and similarity solution:

$$z = x - c\,t, \quad u = h\,(z)$$

into (1) we obtain:

$$g\,h_{zz} + g_h\,\left(h_z\right)^2 + c\,h_z + f = 0 \tag{2}$$

Reduction 2 Generator $\mathbf{v}_3$. Substituting the similarity variable and similarity solution:

$$z = t^{\frac{n-m+1}{2\,m-2}}\,x, \quad u = h\,x^{\frac{2}{n-m+1}}$$

into (1) we obtain:

$$\begin{aligned} &h^n\,(n-m+1)^2\,\left({h_z}^2\,n + h\,h_{zz}\right)\,z^2 + 4\,h^{n+1}\,h_z\,(n+1)\,(n-m+1)\,z \\ &\quad +2\,h^{n+2}\,(n+m+1) + h^{m+1}\,(n-m+1)^2 = 0 \end{aligned}$$

Reduction 3 Generator $\mathbf{v}_4$. Substituting the similarity variable and similarity solution with $n \neq m$:

$$z = t^{\frac{n-m}{2\,m}}\,x, \quad u = \frac{2\,\log(x)}{n-m} + h$$

into (1) we obtain:

$$\begin{aligned} &h_{zz}\,(n-m)^2\,e^{\frac{h\,n^2+h\,m^2}{n-m}}\,z^{\frac{2\,n+m}{n-m}+2} + {h_z}^2\,n\,(n-m)^2\,e^{\frac{h\,n^2+h\,m^2}{n-m}}\,z^{\frac{4\,n-m}{n-m}} \\ &\quad + \left((2\,n+2\,m)\,e^{\frac{h\,n^2+h\,m^2}{n-m}} + \left(n^2 - 2\,m\,n + m^2\right)\,e^{\frac{2\,h\,m\,n}{n-m}}\right)\,z^{\frac{2\,n+m}{n-m}} \\ &\quad + 4\,h_z\,n\,(n-m)\,e^{\frac{h\,n^2+h\,m^2}{n-m}}\,z^{\frac{3\,n}{n-m}} = 0 \end{aligned}$$

Reduction 4 Generator $\mathbf{v}_5$. Substituting the similarity variable and similarity solution with $n \neq 0$:

$$z = x, \quad u = h\,e^{-\frac{c_1\,t}{n}}$$

into (1) we obtain:

$$h^n\,\left(h_z\right)^2\,n + c_2\,h^{n+2} + h^{n+1}\,h_{zz} = 0$$

Reduction 5 Generator $b\mathbf{v}_1 + \mathbf{v}_5$. Substituting the similarity variable and similarity solution:

$$z = \frac{8\,x - 3\,b\,\log t}{8}, \quad u = h\,t^{\frac{3}{4}}$$

into (1) we obtain:

$$-12\,h\,h_{zz} + 16\,\left(h_z\right)^2 - \frac{9\,b\,h^{\frac{7}{3}}\,h_z}{2} + 9\,h^{\frac{10}{3}} - 12\,h^2 = 0$$

Reduction 6 Generator $c\mathbf{v}_3 + \mathbf{v}_6$. Substituting the similarity variable and similarity solution:

$$z = -\frac{4\,c\,e^{-\frac{2\,x}{\sqrt{3}}}}{\sqrt{3}} - \log t, \quad u = h\,e^{-\sqrt{3}\,c\,e^{-\frac{2\,x}{\sqrt{3}}} - \sqrt{3}\,x}$$

into (1) we obtain:

$$27\,h^{\frac{7}{3}}\,h_z\,e^z + 192\,c^2\,h\,h_{zz} - 256\,c^2\,\left(h_z\right)^2 - 96\,c^2\,h\,h_z - 36\,c^2\,h^2 = 0$$

3 Travelling Wave Solutions

We are interested in finding exact travelling wave solutions for Eq. (1) when the diffusion coefficient follows a power law $g(u) = u^m$. From generators $\mathbf{v}_1$ and $\mathbf{v}_2$ we can obtain travelling wave solutions for Eq (1).

To apply the $\frac{G'}{G}$-expansion method to Eq. (2) we suppose that the solutions can be expressed by a polynomial in $\frac{G'}{G}$ in the form

$$h = \sum_{i=0}^{n} a_i \left(\frac{G'}{G}\right)^i, \tag{3}$$

where $G = G(z)$ satisfies the linear second order ODE

$$G''(z) + \omega G'(z) + \zeta G(z) = 0, \tag{4}$$

with a_i, $i = 0, \dots, n$, ω and ζ constants to be determined later and $a_n \neq 0$. The general solutions are well known.

The homogeneous balance between the leading terms provides us with the value of n [6]. Considering the homogeneous balance between h'' and h^2 in (2), we require that $nm + n + 2 = n(m-1) + (n+1)^2 \Rightarrow n = 1$, we can write (3) as

$$h = a_0 + a_1\left(\frac{G'}{G}\right), \quad a_1 \neq 0. \tag{5}$$

From the general solutions of (4), setting without loss of generality $a_0 = a_1 = 1$, we obtain for

$$g(h) = h^m,$$
$$f(h) = \left(h^2 - 2\,h + 2\right)\left(\lambda - 2\,h^{m+1} + 2\,h^m\right) - 2\,h^{m-1}\,\left(h^2 - 2\,h + 2\right)^2\,m$$

the solution

$$h_1(z) = \frac{c_1\,\cos z - c_2\,\sin z}{c_1\,\sin z + c_2\,\cos z} + 1.$$

The corresponding solution for the generalized Fisher equation is

$$u_1 = \frac{c_2\,\sin\left(t\,\lambda - x\right) + c_1\,\cos\left(t\,\lambda - x\right)}{c_2\,\cos\left(t\,\lambda - x\right) - c_1\,\sin\left(t\,\lambda - x\right)} + 1.$$

For

$$g(u) = u^m$$
$$f(u) = (u-2)\,u\,\left(\lambda - u^{m+1}\,m + 2\,u^m\,m - 2\,u^{m+1} + 2\,u^m\right),$$

the corresponding solution for the generalized Fisher equation is

$$u_2 = \frac{c_2\,\sinh\left(x - \lambda t\right) + c_1\,\cosh\left(x - \lambda t\right)}{c_1\,\sinh\left(x - \lambda t\right) + c_2\,\cosh\left(x - \lambda t\right)} + 1. \tag{6}$$

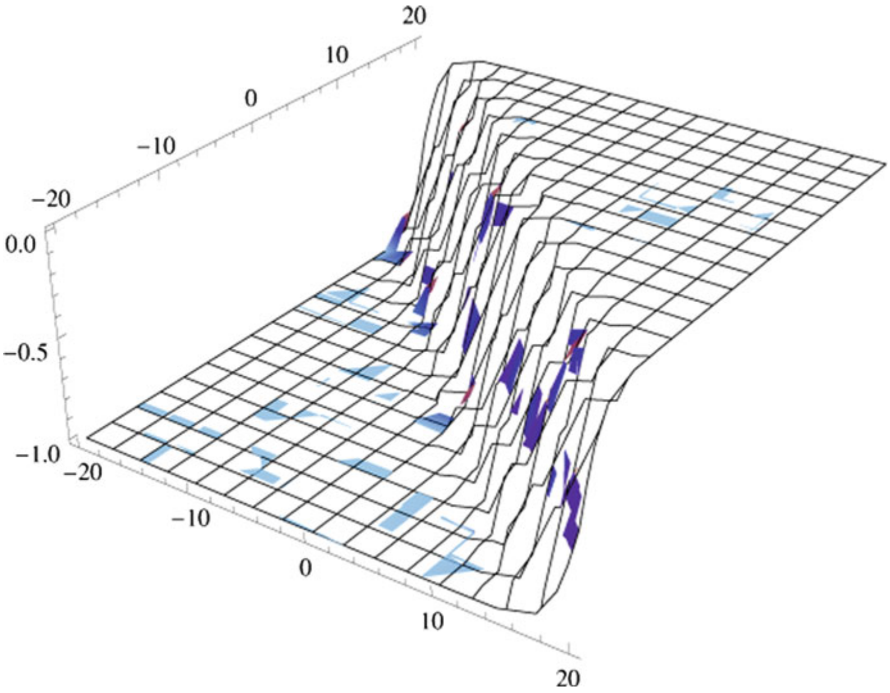


Fig. 1 Kink solution (7), $c_1 = c_2 = \omega = 1$, $\lambda = -1$

Considering the case $\zeta = 0$, $\omega \neq 0$, for

$$\begin{aligned} g(u) &= u^m \\ f(u) &= -u^{m+1}\, m\, \omega^2 - u^{m+1}\, \omega^2 - 2\, u^{m+2}\, m\, \omega - 3\, u^{m+2}\, \omega \\ &\quad + c\, h\, \omega - u^{m+3}\, m - 2\, u^{m+3} + c\, u^2 \end{aligned}$$

$$u(x - \lambda t) = -\frac{c_2\, \omega\, e^{-\omega\,(x-\lambda t)}}{c_2\, e^{-\omega\,(x-\lambda t)} + c_1}, \tag{7}$$

which is a kink solution (Fig. 1).

Acknowledgments The authors acknowledge the financial support from Junta de Andalucía group FQM–201, University of Cadiz.

References

1. Ablowitz, M.J., Zeppetella, A.: Bull. Math. Biol. **41,** 835 (1979)
2. Britton, N.F.: Aggregation and the competitive exclusion principle. J. Theor. Biol. **136,** 57–66 (1989)
3. Bruzon, M.S., Gandarias, M.L.: Symmetry reductions and travelling wave solutions for the Krichever-Novikov equation. Math. Methods Appl. Sci. **35**(8):869–872 (2012)
4. Cherniha, R., Serov, M., Rassokha, I.: Lie symmetries and form-preserving transformations of reaction diffusion convection equations. J. Math. Anal. Appl. **342,** 1363 (2008)
5. Gandarias, M.L., Bruzon, M.S., Rosa, M.: Nonlinear self-adjointness and conservation laws for a generalized Fisher equation. Commun. Nonlinear Sci. Numer. Simul. **18,** 1600–1606 (2013)
6. Kudryashov, N.A.: On “new travelling wave solutions” of the KdV and the KdV–Burgers equations. Commun. Nonl. Sci. Numer. Simulat. **14,** 1891–1900 (2009)
7. Murray, J.D.: Mathematical Biology, 3rd edn. Springer, New York (2002)
8. Olver, P.J.: Applications of Lie Groups to Differential Equations. Springer, Berlin (1986)
9. Wang, M., Li, Xa., Zhang, J.: The (G'/G)-expansion method and traveling wave solutions of nonlinear evolution equations in mathematical physics. Phys. Lett. A **372**, 417–423 (2008). doi: 10.1016/j.physleta.2007.07.051

Numerical Simulation of Potential Maxwell's Equations in the Harmonic Regime

María Teresa González Montesinos and Francisco Ortegón Gallego

Abstract The aim of this work is to perform some numerical experiments for the resolution of a strongly coupled parabolic–elliptic system that describes the heating induction–conduction industrial process of a steel workpiece, whose unknowns are the electric potential, the magnetic vector potential, and the temperature. In order to make the numerical simulations lighter, and taking into account the different time scales between the potentials and the temperature, a new system of nonlinear partial differential equations (PDEs) has been constructed that describes the heating process in the harmonic regime.

1 Introduction

In this work we are concerned with the following nonlinear partial differential equations (PDEs) system

$$-\nabla\cdot(\sigma(\theta)\nabla\varphi) = i\lambda\omega\nabla\cdot(\sigma(\theta)\boldsymbol{A}) + f \;\text{ in } \Omega_T = \Omega\times(0,T), \tag{1}$$

$$\varphi = 0 \;\text{ on } \Gamma_0\times(0,T),\; \frac{\partial\varphi}{\partial n} = -i\lambda\omega\boldsymbol{A}\cdot n \;\text{ on } \Gamma_1\times(0,T), \tag{2}$$

$$i\omega\sigma(\theta)\boldsymbol{A} + L(\boldsymbol{A}) = -\sigma(\theta)\nabla\varphi \;\text{ in } D_T = D\times(0,T), \tag{3}$$

$$\boldsymbol{A} = 0 \;\text{ on } \partial D\times(0,T), \tag{4}$$

$$\rho c_\epsilon\theta_{,t} - \nabla\cdot(\kappa(\theta)\nabla\theta) = \frac{\sigma(\theta)}{2}|i\omega\boldsymbol{A} + \nabla\varphi|^2 + G \;\text{ in } \Omega_T, \tag{5}$$

$$\frac{\partial\theta}{\partial n} = 0 \;\text{ on } \partial\Omega\times(0,T),\; \theta(\cdot,0) = \theta_0 \text{ in } \Omega. \tag{6}$$

F. Ortegón Gallego (✉)
Departamento de Matemáticas, University of Cádiz, Cádiz, Spain
e-mail: francisco.ortegon@uca.es

M. T. González Montesinos
Departamento de Matemática Aplicada I, University of Sevilla, Sevilla, Spain
e-mail: mategon@us.es

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_32

This system describes the heating stage of the induction–conduction industrial procedure applied to a steel workpiece [3–9]. In this framework, $\Omega,\ D \subset \mathbb{R}^3$ are bounded open sets such that $\bar{\Omega} \subset D$, Γ_1 is a relative open set in $\partial\Omega$, $\Gamma_0 = \partial\Omega \setminus \Gamma_1$; also these sets and boundaries are supposed to be smooth enough. Also, $T > 0$ is the final time of the heating process, and the unknowns are the electric potential, φ, the magnetic vector potential, $\boldsymbol{A}$, and the temperature, θ; σ and κ are the electric and thermal conductivities, respectively, ω is the angular frequency, θ_0 the initial temperature, i stands for the imaginary unit, $\lambda \in [0, 1]$ is a parameter, G is a source term coming from phase transitions of steel and mechanical deformations, ρ is the density, and c_ϵ is the specific heat at constant pressure.

Problems (1)–(6) are referred to as the harmonic regime [1, 2, 5, 6]. In this way, if $\phi : \Omega \times [0, T] \mapsto \mathbb{R}$ and $\mathcal{A} : \Omega \times [0, T] \mapsto \mathbb{R}$ are the electric and the magnetic vector potentials in the original problem, respectively, we may write $\phi = \mathrm{Re}[\exp(i\omega t)\varphi(x, t)]$ and $\mathcal{A} = \mathrm{Re}[\exp(i\omega t)\boldsymbol{A}(x, t)]$, where φ and $\boldsymbol{A}$ are complex–valued fields, and they are called in the same way by abuse of language. Finally, $L \in \mathcal{L}(\boldsymbol{W}, \boldsymbol{W}')$ is some elliptic operator, being $\boldsymbol{W}$ a suitable Hilbert space.

In the original model is $\lambda = 1$, but in most numerical simulations the value $\lambda = 0$ is taken which yields to an enormous reduction of computational cost. In [7], the authors have shown the existence of a weak solution to (1)–(6) in the range $0 \leq \lambda < 1 - \frac{1}{\omega}$, so that $\lambda = 1$ is not attainable!

We have carried out some numerical experiments for the resolution of the linear system φ–$\boldsymbol{A}$ for a given temperature for different values of the parameter λ. We have used a Crank–Nicolson like iterative scheme. The numerical results show a strong relation between λ and the rate of convergence of this scheme: the closer the value of λ to the critical value $1 - 1/\omega$, the more number of iterations are needed. These results have been obtained using the FreeFem++ software (see [8]).

2 Notation and Assumptions on Data

Let V and $\boldsymbol{W}$ be Hilbert spaces such that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$, where Poincaré's inequality is fulfilled, and $H_0^1(D)^3 \subset \boldsymbol{W} \subset H^1(D)^3$.

The following hypotheses are assumed on data:

(H.1) $\sigma,\ \kappa : \Omega \times \mathbb{R} \mapsto \mathbb{R}$ are Carathéodory functions and there exist some constant values $\sigma_1,\ \sigma_2\ \kappa_1,\ \kappa_2$ such that

$$0 < \sigma_1 \leq \sigma(x, s) \leq \sigma_2, \quad 0 < \kappa_1 \leq \kappa(x, s) \leq \kappa_2,$$

almost everywhere $x \in \Omega$ and for all $s \in \mathbb{R}$.

(H.2) $L \in \mathcal{L}(\boldsymbol{W}, \boldsymbol{W}')$ is such that, for some constant value $\alpha > 0$,

$$\langle L(\bar{w}), \bar{w}\rangle_{W',\boldsymbol{W}} \geq \alpha \|w\|^2_{H^1(D)^3}, \text{for all } \bar{w} \in \boldsymbol{W}. \tag{7}$$

(H.3) $\lambda \in \left[0, 1 - \frac{1}{\omega}\right)$.

(H.4) $f \in L^2(V')$.
(H.5) $G \in L^1(\Omega_T)$ and $\theta_0 \in L^1(\Omega)$.

Also, the space $\boldsymbol{W}$ will be defined depending on the linear operator L described in (H.2). For instance, we may consider

1. $\boldsymbol{W} = H_0^1(D)^3$ and L is given by

$$L(\boldsymbol{w}) = \nabla\times\left(\frac{1}{\mu}\nabla\times\boldsymbol{w}\right) - \delta\nabla\,(\nabla\cdot\boldsymbol{w}), \tag{8}$$

   where μ is the magnetic permeability and $\delta > 0$ is a small parameter, or

$$L(\boldsymbol{w}) = -\Delta\boldsymbol{w}. \tag{9}$$

2. $\boldsymbol{W} = \{\boldsymbol{w} \in H^1(D)^3 \,/\, \nabla\cdot\boldsymbol{w} = 0 \text{ in } D,\ \boldsymbol{w}\times\mathbf{n} = 0 \text{ on } \partial D\}$ with $\partial D \in C^{1,1}$ and

$$L(\boldsymbol{w}) = \nabla\times\left(\frac{1}{\mu}\nabla\times\boldsymbol{w}\right). \tag{10}$$

3 An Existence Result

Now we state an existence result related to system (1)–(6) (see [7]).

Theorem 1 *Under hypotheses (H.1)–(H.5) problem (1)–(6) has a weak solution* $(\varphi, \boldsymbol{A}, \theta)$. *Moreover, for any* $\lambda \in [0, 1 - \frac{1}{\omega})$, *there exists a constant* C_λ *such that*

$$\int_\Omega \sigma(\theta)|\nabla\varphi|^2 + \int_D \sigma(\theta)|\omega\boldsymbol{A}|^2 \leq C_\lambda,$$

where

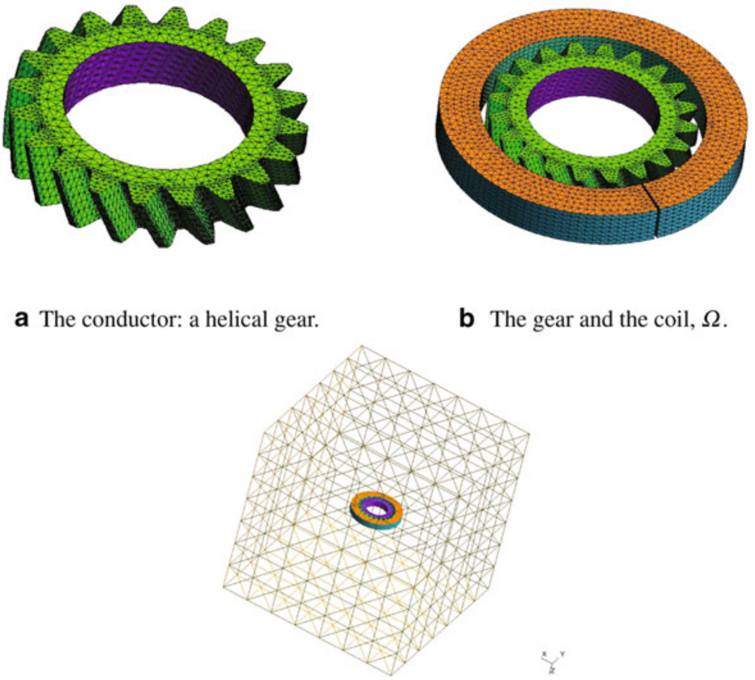
$$\lim_{\lambda\to(1-1/\omega)^-} C_\lambda = +\infty.$$

4 Numerical Scheme and Some Results

In what follows, the elliptic operator L is given by (8), and θ is a fixed temperature. In order to analyze the sensitivity of $(\varphi, \boldsymbol{A})$ with respect to λ as $\lambda \to 1 - 1/\omega$, we have performed some numerical simulations for the resolution of the linear system (1)–(4) for some values of λ. We consider a Crank–Nicolson like scheme to approximate the solution $(\varphi, \boldsymbol{A})$ as follows.

Initialization. The functions φ^0 and $\boldsymbol{A}^0$ are given by the solution of the respective variational equations

$$\int_\Omega \sigma\nabla\varphi^0\nabla\bar{v} = \langle f, \bar{v}\rangle_{V',V}, \quad \text{for all } v \in V, \tag{11}$$



a The conductor: a helical gear.

b The gear and the coil, Ω.

c The whole domain, D.

Fig. 1 Description of the domains considered in the numerical resolution. In this setting, Ω is the set of conductors, that is, the gear (steel) and the coil (copper). Here, the boundary Γ_0 is the two opposite square faces in the coil. On the other hand, the domain D, where is defined the magnetic vector potential $\boldsymbol{A}$ is the *big box* containing the workpiece together with the coil

$$i\omega \int_{\Omega} \sigma \boldsymbol{A}^0 \bar{w} + \left\langle L\left(\boldsymbol{A}^0\right), \bar{w}\right\rangle_{\boldsymbol{W}',\boldsymbol{W}} = -\int_{\Omega} \sigma \nabla \varphi^0 \bar{w}, \quad \text{for all } w \in \boldsymbol{W}. \tag{12}$$

For $n \geq 0$: Assume $(\varphi^n, \boldsymbol{A}^n)$ is known and compute $\tilde{\varphi}^{n+1}$ then $\tilde{\boldsymbol{A}}^{n+1}$ according to

$$\int_{\Omega} \frac{\tilde{\varphi}^{n+1} - \varphi^n}{k/2} \bar{v} + \int_{\Omega} \sigma \nabla \tilde{\varphi}^{n+1} \nabla \bar{v} = -i\omega\lambda \int_{\Omega} \sigma \boldsymbol{A}^n \nabla \bar{v} + \langle f, \bar{v} \rangle_{V',V}, \quad v \in V, \tag{13}$$

$$\int_{D} \frac{\tilde{\boldsymbol{A}}^{n+1} - \boldsymbol{A}^n}{k/2} \bar{w} + i\omega \int_{D} \sigma \tilde{\boldsymbol{A}}^{n+1} \bar{w} + \left\langle L\left(\tilde{\boldsymbol{A}}^{n+1}\right), \bar{w}\right\rangle_{\boldsymbol{W}',\boldsymbol{W}} = -\int_{\Omega} \sigma \nabla \tilde{\varphi}^{n+1} \bar{w}, \quad \text{for all } w \in \boldsymbol{W}. \tag{14}$$

and the new iteration is given by $\varphi^{n+1} = 2\tilde{\varphi}^{n+1} - \varphi^n$, $\boldsymbol{A}^{n+1} = 2\tilde{\boldsymbol{A}}^{n+1} - \boldsymbol{A}^n$. If $\|\varphi^{n+1} - \varphi^n\| < \epsilon$, STOP: We keep $(\varphi^{n+1}, \boldsymbol{A}^{n+1})$ as an approximation to $(\varphi, \boldsymbol{A})$.

On the other hand, we focus our attention in a specific domain, namely, a helical gear, as it is shown in Fig. 1. This particular setting is usually found in the industrial

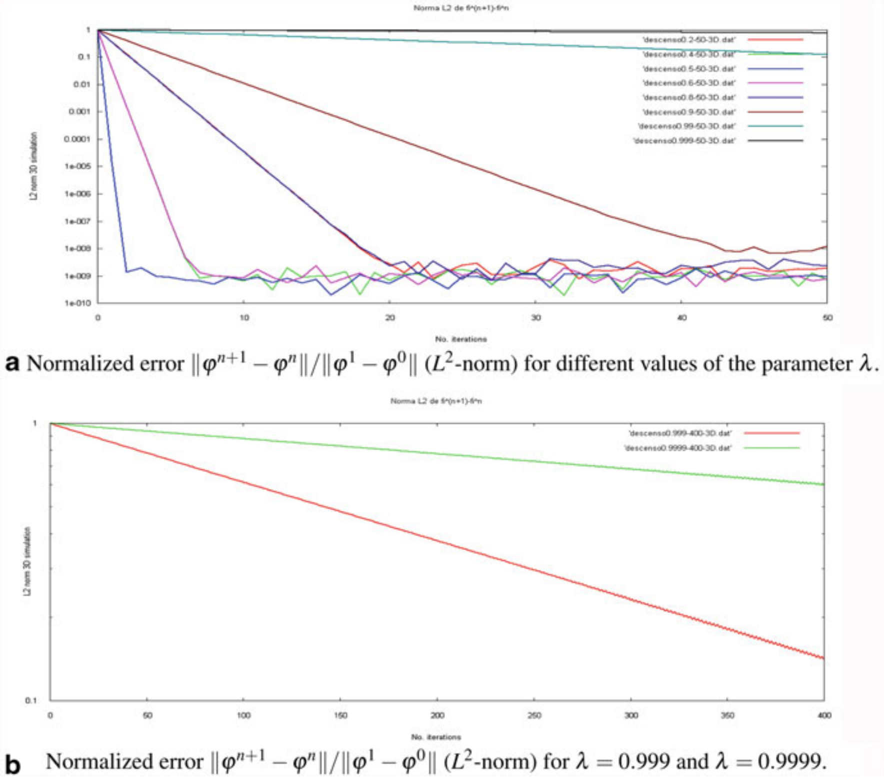
Table 1 Tetrahedralization data of the domains considered in the numerical simulations

	Gear	Coil	Box
No. vertices	5317	6924	26,468
No. edges	8560	6528	27,432
No. tetrahedra	18,881	30,937	159,809

heating process by induction [10]. Here, the frequency is 900 Hz, $\omega = 2\pi \times 900 = 5,654.88\ldots$ and thus, $1 - 1/\omega = 0.999823\ldots$. The numerical resolution of the variational formulations (11)–(14) has been obtained by means of the finite element method using a P_1-Lagrange approximation for both φ and the three components of $\boldsymbol{A}$. Table 1 gives some data of the tetrahedralization of the domains considered in this numerical simulation. Figure 2a shows the strong sensitivity of the rate of convergence of the approximate solutions with respect to the parameter λ: The closer the value of λ to $1 - 1/\omega$ the more iterations are needed to achieve convergence. In Fig. 2b, we consider the values $\lambda = 0.999$ and $\lambda = 0.9999$. Here we show the normalized error for 400 iterations of the algorithm (13)–(14); notice that at about 200 iterations, this algorithm becomes numerically unstable so that convergence is not guaranteed.

5 Conclusions

We have carried out some numerical experiments in order to approximate the solution to a nonlinear coupled system of PDEs describing the heating industrial process of a steel workpiece by induction. This system depends on a parameter λ for which the mathematical analysis assures the existence of a weak solution for $\lambda \in [0, 1 - 1/\omega)$ such that the associated energy blows up as $\lambda \to (1 - 1/\omega)^-$. In accordance with these theoretical results, the numerical experiments show a strong sensitivity with respect to λ. Indeed, for values of λ close to the upper bound $1 - 1/\omega$ the numerical algorithm becomes unstable since the computed normalized difference of two consecutive iterations develops fluctuations.



a Normalized error $\|\varphi^{n+1} - \varphi^n\| / \|\varphi^1 - \varphi^0\|$ (L^2-norm) for different values of the parameter λ.

b Normalized error $\|\varphi^{n+1} - \varphi^n\| / \|\varphi^1 - \varphi^0\|$ (L^2-norm) for $\lambda = 0.999$ and $\lambda = 0.9999$.

Fig. 2 **a** Normalized difference of two consecutive iterations of the sequence (φ^n) according to the numerical scheme (11)–(14) for some values of the parameter λ, namely $\lambda = 0.2$, 0.4, 0.5, 0.6, 0.8, 0.9, 0.99, and 0.999. We carried out 50 iterations in every case. The closer the value of λ to $1 - 1/\omega$ the worse the rate of convergence. **b** The same normalized difference for $\lambda = 0.999$ and $\lambda = 0.9999$. Here we have carried out 400 iterations without attaining convergence. At about $n = 200$ fluctuations seem to develop. Notice that the value $\lambda = 0.9999$ is outside the interval $[0, 1 - 1/\omega)$ so that neither the existence of a solution from Theorem 1 nor the convergence of the scheme (11)–(14) are assured

Acknowledgment This research was partially supported by Ministerio de Economía y Competitividad under grant MTM2010-16401 with the participation of FEDER, and Consejería de Educación y Ciencia de la Junta de Andalucía, research group FQM-315. The authors wish to thank the referees and Editorial comments for they have helped us to improve the presentation of this chapter.

References

1. Bossavit, A.: Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements. Academic Press, San Diego (1997)
2. Clain, S.: Analyse mathématique et numérique d'un modèle de chauffage par induction. Ph.D. Thesis, N. 1240, Laussane, EPFL (1994)

3. Díaz Moreno, J.M., García Vázquez, C., González Montesinos, M.T., Ortegón Gallego, F.: Analysis and numerical simulation of an induction–conduction model arising in steel heat treating. J. Comp. Appl. Math. **236**, 3007–3015 (2012)
4. Díaz Moreno, J.M., García Vázquez, C., González Montesinos, M.T., Ortegón Gallego, F., Viglialoro, G.: Mathematical modeling of heat treatment for a steering rack including mechanical effects. J. Numer. Math. **20**(3–4), 215–231 (2012)
5. González Montesinos, M.T., Ortegón Gallego, F.: Analysis of a nonuniformly elliptic and nonlinear coupled parabolic–elliptic system arising in steel hardening. Int. J. Comput. Math. (2013), http://dx.doi.org/10.1080/00207160.2013.771837
6. González Montesinos, M.T., Ortegón Gallego, F.: On an induction–conduction PDEs system in harmonic regime. Nonlinear Anal.: Real World Appl. **15**, 58–66 (2014), http://dx.doi.org/10.1016/j.nonrwa.2013.05.006
7. González Montesinos, M.T., Ortegón Gallego, F.: On the existence of a weak solution to a strongly coupled system in harmonic regime arising in steel hardening, (to appear)
8. Hecht, F., Pironneau, O., Le Hyaric, A., Ohtsuda, K.: FreeFem++ (Third Edition, Version 3.19). Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, 2012. http://www.freefem.org/ff++/
9. Hömberg, D.: A mathematical model for induction hardening including mechanical effects. Nonlinear Anal. Real World Appl. **5**, 55–90 (2004)
10. Otto, F.J., Herring, D.H.: Gear heat treatment. Part I. Heat Treating Progress, 1–5 (2002)

Supply Chain Flexibility Metrics Evaluation

Mujde Erol Genevois, Ugur Gure and Kaya Ocakoglu

Abstract The markets in which manufacturers and service firms compete are increasingly influenced by intense foreign competition, rapid technological change, and shorter product life cycles. In this new scenario, flexibility may be one of the most important capabilities needed for firms to achieve competitive advantage. The possible behaviors of the company to the problems it faces are called levers of flexibilities. In a supply chain, the flexibility of one entity is highly dependent on the flexibility of upstream entities. It is a natural area for metrics. A metric is a standard of measurement of performance and gives the basis on which to evaluate the performance of processes in the supply chain. Thus, the purpose of the study is to determine and evaluate the supply chain flexibility levers in order to calculate the benefit of preferring a flexibility lever to another one. The analytic network process (ANP) technique is used for prioritizing evaluated flexibility levers. We are handling the automotive sector for the study.

1 Introduction

Automobile manufacturers today compete in an increasingly global environment. An important part of the equation for competing in today's automotive industry is flexibility. Cadences are tightening to respond to market demands, but manufacturers need to be even more flexible than that. Inflexibility equals lost opportunities. Today's manufacturing line needs to be flexible and agile, which has come about through configurability, distributed control and plug-and-play capabilities. Obviously, the exibility is deployed more often in segments with higher proportion of exible competitors.

M. E. Genevois (✉) · U. Gure · K. Ocakoglu
Industrial Engineering Department, Galatasaray University, Istanbul, Turkey
e-mail: merol@gsu.edu.tr

U. Gure
e-mail: uur.gure@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_33

This study is focused on passenger cars and on segments which are most preferred by customers according to sales numbers. Only four segments will be investigated: A—Basic, B—Small, C—Lower Medium, D—Upper Medium. For a clear understanding, Ford Ka is an example to A class, Volkswagen Polo is an example to B class, Toyota Auris is an example to C class, and BMW 3 series is an example to D class cars.

In this chapter, customer expectations satisfaction via adapting automotive industry flexibility will be studied. Flexibility is defined as the capacity of responding against uncertainties created by various causes in the environment. Possible actions to ensure flexibility are called as levers of flexibility and their performance evaluation tools are called metrics of flexibility. First, automotive industry will be briefly presented via its three actors expectations; supplier, producer, and customer. Second, the concept of flexibility and its importance will be investigated. Third, the methodology including the analytic network process (ANP) technique for prioritizing evaluated flexibility levers by a group of experts will be presented. Finally, the outcome will be discussed according to the results, the metrics to evaluate the system performance will be defined, and possible investments will be proposed.

2 Automotive Sector's Expectations and Related Metrics

Every supply chain has three aspects which are customer, producer, and supplier. In automotive sector, all these three aspects have distinctive and also some common expectations such as cost-minimizing, efficiency, technological advance, sustainability, environmentally friendly production, endurance, reliability, etc.

Customer expectations are considered as customization, high responsiveness, delivery reliability, right quality, and after sales services. Manufacturing firms aim to achieve the highest levels of performance along areas such as quality, flexibility, delivery, and costs [1]. In this study main producer and supplier expectations are considered as process optimization, supply reliability, loyal customer, minimum consumption of resources, and effective risk management.

Metrics are tools for measuring performance. Supply chain operation reference (SCOR) model provides a measure of supply chain performance by dividing it into four parts: plan, source, make, and deliver [2]. According to the literature survey and experts feedback, suitable metrics for ASCI are: forecast accuracy, in-stock availability, perfect order fulfillment, materials quality, weekly/monthly plan keeping, production lead time track, days of inventory track, capacity utilization, output/input ratio, labor performance, and vendor lead time track [3].

3 Flexibility Management in Automotive Sector

Investment channels of the automotive sector are broad and multinational. Also, automotive sector has a high ratio of supply chain cost to revenue. Various drivers should cooperate to ensure efficiency in a supply chain. A key dimension of supply chain performance is flexibility, i.e., the ability to be adapted to internal and external capabilities or a reaction to environmental uncertainty [4].

In literature, it is easy to find various previous studies on flexibility in automotive sector. Barad and Sapir in 2003 [5] studied logistics flexibility. They presented flexibility types and quantitatively investigated one of the dimensions. Sanchez and Perez in 2005 [6] studied supply chain flexibility and firm performance. They clearly defined supply chain flexibility and its subdivisions. Erol Genevois and Gurbuz in 2009 [7] studied flexibility in automotive sector and utilized fuzzy hierarchical process method to determine flexibility levers which can best meet the customer satisfaction.

To prevent confrontations between flexibility levers, we grouped levers under five main parts. These are a supply chain's vital components: mix, volume, delivery, quick-design change, and adaptation levers.

We define mix flexibility as actions against uncertainty as to which products customers will accept leads to the strategic objective of product diversity. Mix flexibility spans modification flexibility (MF) which allows a manufacturing process to implement minor design changes in a given product, decision-making flexibility (DMF) which is an intangible lever ensured by intelligent management of the system. According to us and experts, DMF is the core of the effective management, planning/scheduling flexibility (P/SF), and sequencing flexibility (SF).

Volume flexibility permits increases or decreases in the aggregate production level. It spans labor flexibility (LF), material flexibility which is the ability of the manufacturing function to handle unexpected variations in inputs, DMF, P/SF, SF, and routing flexibility (RF) which is the capability of processing a part through varying routes, or in other words by using alternative machines [8].

Delivery flexibility permits to construct systems that ease to meet true demand in true place and at true time. Delivery flexibility spans transport/shipping flexibility (T/SF), access flexibility (AF) which is demanded for responding customer needs agile as possible, DMF, P/SF, and SF.

Quick design change flexibility is required to ensure company's continuous competitiveness in the market. Banking flexibility is also possible [9]. Quick design change flexibility spans launch flexibility (LchF), design development flexibility (D/DF), changeover flexibility (CF), DMF, and job design flexibility (JDF).

The capability of a manufacturing system that enables it to adapt rapidly and inexpensively to changes in its internal and external operating environment is called adaptation flexibility [10]. Adaptation flexibility spans process/technology flexibility (P/TF), machine/equipment flexibility (M/EF), material flexibility (MatF), employee's willingness to change flexibility (EWF), managerial perception change flexibility (MPCF), LF, layout flexibility (LayF), expansion flexibility (EF), financial resources flexibility (FRF), and organizational structure flexibility (OSF).

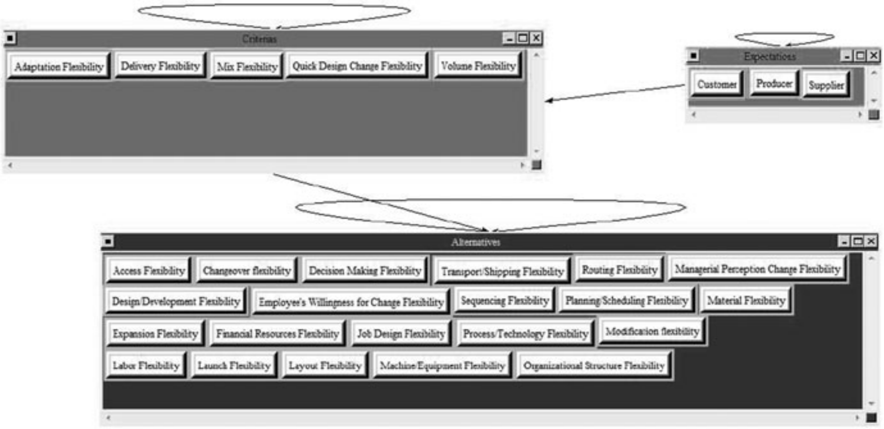

Criterias
Adaptation Flexibility
Delivery Flexibility
Mix Flexibility
Quick Design Change Flexibility
Volume Flexibility
Expectations
Customer
Producer
Supplier
Alternatives
Access Flexibility
Changeover flexibility
Decision Making Flexibility
Transport/Shipping Flexibility
Routing Flexibility
Managerial Perception Change Flexibility
Design/Development Flexibility
Employee's Willingness for Change Flexibility
Sequencing Flexibility
Planning/Scheduling Flexibility
Material Flexibility
Expansion Flexibility
Financial Resources Flexibility
Job Design Flexibility
Process/Technology Flexibility
Modification flexibility
Labor Flexibility
Launch Flexibility
Layout Flexibility
Machine/Equipment Flexibility
Organizational Structure Flexibility

Fig. 1 1 The analytic network process (ANP) network scheme of the decision problem

4 Methodology

ANP is a multi criteria decision making tool considered to be an extension of analytic hierarchy process (AHP) [11]. Whereas AHP models a decision making framework using a unidirectional hierarchical relationship among decision levels, ANP allows for more complex interrelationships among the decision levels and components, like a network [12].

Step 1: The first step is defining our decision problem and then model to be evaluated is constructed. The main objective of the problem is to evaluate the satisfaction degree of automotive sector actors' expectations via attributed flexibility levers. This model has three clusters and their nodes are: expectations (supplier expectations, producer expectations, and customer expectations), flexibility types (mix flexibility, volume flexibility, quick design change flexibility, delivery flexibility, adaptation flexibility) and flexibility levers (MF, DMF, P/SF, SF, LF, MatF, RF, T/SF, AF, LchF, D/DF, CF, JDF, P/TF, M/EF, EWF, MPCF, LayF, EF, FRF, OSF).

Step 2: Given this model, the relevant criteria and alternatives are structured in the form of a simple network by the decision makers. Interdependencies are represented by the arrows among the clusters (outer dependence) and a looped arc within the same cluster (inner dependence). The direction of the arc signifies dependence. Arcs emanate from a controlling attribute to other attributes that may influence it. All the relations among criteria and sub-criteria, and the network of the model can be seen in Fig. 1.

Table 1 Final results

Flexibility levers	Normal
Decision making flexibility	0.1604
Planning/scheduling flexibility	0.1526
Material flexibility	0.1084
Financial resources flexibility	0.0983
Design/development flexibility	0.0926
Transport/shipping flexibility	0.0762
Changeover flexibility	0.0723
Process/technology flexibility	0.0647
Sequencing flexibility	0.0405
Expansion flexibility	0,0895
Others	0,0762

Step 3: In this step of the ANP methodology, comparison sets between clusters and elements are set. To build the comparison matrices, clusters and their elements are compared with respect to a control criterion. To reflect interdependencies in this simple network model, pairwise comparisons among all the clusters/elements/alternatives are performed and these relationships are evaluated. As for the evaluation of the alternatives and criteria, the fundamental comparison scale (1 to 9) is used.

The ANP method is able to handle interdependencies among elements through the calculation of composite weights as developed in a supermatrix. After completing all the pairwise comparisons, the derived priorities of the unweighted supermatrix are obtained for each control criterion. Then, using the cluster weights matrix, the priorities of all factors in each cluster are weighted. The weighted supermatrix, each of whose columns sums to one, is known as a column stochastic matrix. The weighted supermatrix is then raised to limit powers to obtain the final priorities of all elements in the limit matrix. Then the results are synthesized through addition for the entire control criterion. These synthesized results of the priorities are normalized to select the highest priority alternative. The supermatrix and its powers are the fundamental tools needed to lay out the functions of the ANP [13].

Step 4–5–6: The experts' opinions are used to fill in the pairwise comparison matrices for all clusters and then the supermatrix is built according to these pairwise comparison matrices by using the Super Decisions software. Pairwise comparisons tables are completed in consensus by five experts who work in automotive industry.

Step 7: Given the comparison matrices, the Super Decisions software computed the unweighted, weighted, and limit supermatrices. The synthesized results and the priorities are provided.

Step 8: Finally, the first ranking flexibility levers are synthesized and are shown in Table 1. DMF, P/SF, and Mat F have the highest rankings in our final result.

5 Conclusion

In this study, a decision-making model, based on ANP is developed. The needs for DMF, P/SF, and MatF are highly important in the automotive sector. DMF has 16% importance in all levers because its the key factor for quick response to uncertainties and satisfies expectations. It must be ensured with metrics such as forecast accuracy, inventory turnover, and planning cycle time analysis. P/SF has 15% importance. This lever is very important for mix and delivery flexibilities which are essential for satisfying customer and producer expectations. Weekly/daily plan keeping analysis, production lead time track, capacity on time shipment ratio, and on time delivery ratio metrics can be utilized for measuring P/SF. An average car has 12,000 different parts. Thats why MatF has a crucial role in a flexible supply chain. Material quality, input/output ratio are possible metrics to measure this flexibility. For the future works, the study will be developed with a metrics quantification dimension.

Acknowledgement This research has been financially supported by Galatasaray University Research Fund. (Research no: 12.402.008)

References

1. Silveira, G., Borenstein, D., Fogliatto, F.: Mass customization: literature review and research directions. Int. J. Prod. Econ. **72,** 1–13 (2001)
2. Stewart, G.: SCOR: the first cross-industry framework for integrated supply-chain management. Logist. Inf. Manage. **10**(2), 62–67 (1997)
3. Chae, B.: Developing key performance indicators for supply chain: an industry perspective. Supply Chain Manage. Int. J. **14**(6), 422–428 (2009)
4. Vickery, S., Calantone, R., Droge, C.: Supply chain flexibility: an empirical study. J. Supply Chain Manage. **35**(2), 16–24 Summer (1999)
5. Barad, M., Sapir, E.: Flexibility in logistic systems? Modeling and performance evaluation. Int. J. Prod. Econ. **85,** 155–170 (2003)
6. Sanchez, A., Pérez, M.: Supply chain flexibility and firm performance—A conceptual model and empirical study in the automotive industry. Int. J. Oper.& Prod. Manage. **25**(7), 681–700 (2005)
7. Genevois, M., Gürbüz, T.: Finding the best flexibility strategies by using an integrated method of FAHP and QFD. IFSA-EUSFLAT (2009)
8. Gupta, Y., Goyal, S.: Flexibility of manufacturing systems: concepts and measurement. Eur. J. Oper. Res. **43**(2), 119–135 (1989)
9. Kramer, A., Kramer, J.: Flexibility of delivery frequency in logistics competition. Working paper, available on social science research network (2010)
10. Swamidass, M.: Encylopedia of Production and Manufacturing Management. Kluwer, Dordrecht (2000)
11. Saaty, T.L.: The Analytical Hierarchy Process. McGraw Hill, New York (1981)
12. Sarkis, J.: Evaluating environmentally conscious business practices. Eur. J. Oper. Res. **107**(1), 159–174 (1998)
13. Saaty, T.L.: Decision Making with Dependence and Feedback: The Analytic Network Process, 2nd edn. RWS Publications, Pittsburgh (2001)

Estimation of Abundance from a Correlated Binary Time Series

Julie Horrocks, Matthew Rueffer, David Hamilton and Sarah Wong

Abstract In the face of increasing extinction rates, it is vital to have estimates of relative and absolute species abundance and their relationship to important factors. For species that live in the oceans or large lakes, this can be a difficult task. Here, we present a method for estimating absolute abundance from a single binary acoustic time series. The dependence in the series is exploited to allow the estimation of abundance when some animals remain hidden, and in the face of uncertainty about the range over which sounds carry. Simulations show that the method works well, even when some assumptions are violated. The method is illustrated using data on sperm whales in the Sargasso Sea.

1 Introduction

For the purposes of conservation biology, estimating the abundance of wild animals is critical to monitor the status of populations. Determining the relationship between abundance and various external factors is particularly important for the development of conservation and management strategies, especially in the face of increasing pressures from habitat fragmentation, environmental degradation, and global warming.

J. Horrocks (✉) · M. Rueffer
University of Guelph, Guelph, ON, Canada
e-mail: jhorrock@uoguelph.ca

M. Rueffer
e-mail: mrueffer@uoguelph.ca

D. Hamilton · S. Wong
Dalhousie University, Halifax, NS, Canada
e-mail: david.hamilton@dal.ca

S. Wong
e-mail: snpwong@dal.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_34

2 General Methods for Estimating Abundance

Methods for estimating abundance or relative abundance can be categorized as follows (see [2]):

- Random sampling of areas: divide a closed area into sectors, take a random sample of sectors, and count the number of animals in each sector. If all animals which are present are detected, an accurate estimate of density and hence absolute abundance of animals in the area can be made.
- Distance sampling: record the number of animals detected along a transect within a strip of a given width and estimate the distance from the transect for each individual. This method operates under the assumption that animals close to the transect are perfectly detected, while detectability decreases as distance from the transect increases. A detectability function is estimated from the data, and density and absolute abundance can be estimated.
- Effort methods: these methods operate under the assumption that the more effort one puts into looking for animals, the more animals will be detected. They are generally used to estimate relative abundance.
- Capture recapture: in a closed population, a random sample of animals is selected, marked, and released back into the population and allowed to mix. A second sample is taken, which will presumably include some of the previously marked individuals. Assuming that the proportion of marked animals in the second sample equals the proportion of marked animals in the whole population, the size of the population can be estimated. Various modifications of this method exist, allowing for open populations, more than two samples, etc. These methods can be used to estimate absolute abundance.

Here, we will describe a particular method for estimating abundance of aquatic animals from acoustic data which could be considered a capture recapture method.

3 Estimating Abundance from Correlated Binary Acoustic Data

The survey method described here was developed by Whitehead [3], who used it to estimate abundance of whales from acoustic data, using a method-of-moments estimator. This was generalized to maximum likelihood by Horrocks et al. [1], as we will now describe.

Imagine that a researcher sails along a transect and listens for whales at regular intervals. While an expert can distinguish sperm whale vocalizations from other species, it is not possible to discern which individual is vocalizing or even how many individuals are vocalizing. Therefore at each interval, the researcher records either a 1 (if whales were heard) or a 0 (if no whales were heard).

In order to estimate absolute abundance from such imprecise data, it is necessary to make some assumptions. We will assume that, if a whale is present within a circle of radius r with the researcher at the center and it vocalizes, then it may be detected by the researcher, but if it is outside this “listening circle” it cannot be detected. As the researcher moves along the transect, a series of listening circles is defined. The distance between the centroids of the listening circles is assumed to be known and constant across the survey. Further assumptions are as follows:

- Individuals (or groups) are distributed according to a Poisson process with constant intensity α. This amounts to assuming that individuals are uniformly distributed in space.
- Individuals move slowly with respect to the boat; i.e., do not follow or avoid it.
- Range of detection r is constant during the survey.
- Probability of vocalization/detection μ is constant during the survey.
- Vocalizations are independent between and within individuals (or groups). We are assuming that individuals may vocalize multiple times.

Violations of these assumptions were examined by simulation. There are three parameters which must be estimated (α, μ, and r) but our main interest is in α.

Our method exploits dependence in the data to estimate the three parameters. If no listening circles overlapped, we would have independent binary data, and could estimate only one parameter, namely, the proportion of listening circles in which whales were heard. Our method requires that listening circles overlap so that a single whale can be heard more than once. Thus, the method can be considered as a sort of capture recapture method, where a “capture” corresponds to hearing a whale and a “recapture” corresponds to hearing the same whale more than once. The overlapping of the listening circles creates dependence in the data and allows us to estimate more than one parameter. As observations are not independent, the likelihood is constructed as a product of conditional probabilities. We used a second-order Markov approximation to the likelihood:

$$L(\alpha,\mu,r) = \prod_{t=3}^{N} \left[P(Y_t|Y_{t-1},Y_{t-2})\right] P(Y_2,Y_1),$$

where Y_t is the binary outcome at time t, $P(Y_t|Y_{t-1},Y_{t-2})$ is the probability of the outcome at time t given the outcomes at times $t-1$ and $t-2$, $P(Y_2,Y_1)$ is the joint probability of the outcomes at times 1 and 2, and N is the number of listening circles.

Since the parameters α, μ, and r are assumed to be constant over time, the probabilities are also independent of time and the likelihood simplifies to

$$L(\alpha,\mu,r) = \left(\frac{P_{000}}{P_{00}}\right)^{n_{000}} \left(\frac{P_{100}}{P_{10}}\right)^{n_{100}} \cdots \left(\frac{P_{110}}{P_{11}}\right)^{n_{110}} \left(\frac{P_{011}}{P_{01}}\right)^{n_{011}} \left(\frac{P_{111}}{P_{11}}\right)^{n_{111}}$$
$$P_{00}^{(1-y_1)(1-y_2)} P_{01}^{(1-y_1)(y_2)} P_{10}^{(y_1)(1-y_2)} P_{11}^{y_1 y_2},$$

where $P_{ij} = P(Y_{t-1} = i, Y_t = k)$ and $P_{ijk} = P(Y_{t-2} = i, Y_{t-1} = j, Y_t = k)$ for $t = 3, 4, \ldots, N$, and n_{ijk} equals the number of triplets where $y_{t-2} = i$, $y_{t-1} = j$, and $y_t = k$. Here i, j, k can take values 0 or 1. Expressions for the P_{ijk} depend on the geometry of overlapping parts of circles, and are given in Horrocks et al. [1].

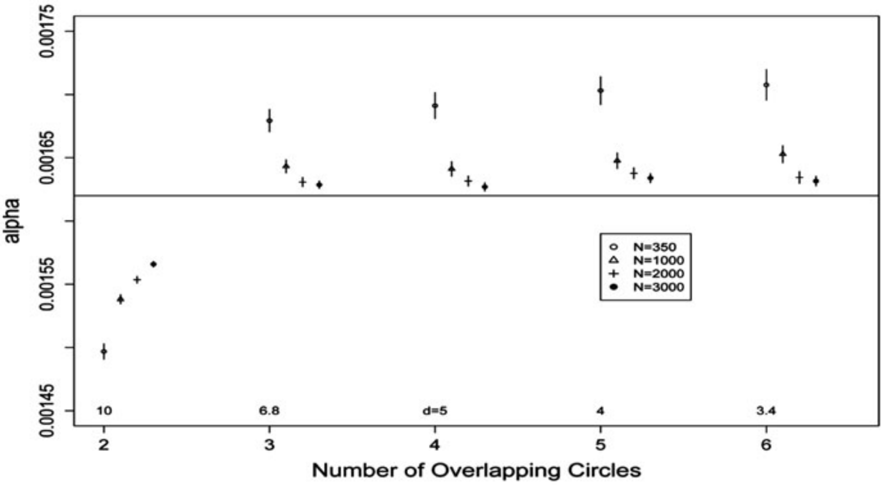


Fig. 1 Bias in α when assumptions are satisfied, for varying values of d, the distance between listening circles, which induces varying numbers of overlapping circles. The true value of $\alpha(0.00162)$ is shown as a *horizontal line*

We conducted extensive simulations to study the performance of the method. First, we examined the effect of different spacings between the listening circles, when all assumptions are satisfied. We generated the positions of individuals over a rectangular area from a Poisson process with density α. A sequence of N listening circles ($t = 1, 2, \ldots N$) with radius r and distance between centroids of d were then generated. If an individual fell within the tth listening circle, we assumed it was detected with probability μ, in which case y_t was set to 1, else $y_t = 0$. We simulated a total of 10,000 data sets.

Figure 1 shows results when $\alpha = 0.00162$, $\mu = 0.9$, and $r = 10$. Points with the same value of d have been shifted slightly in the horizontal direction for clarity.

The method performs well even when the number of overlapping circles is not 3. For $N = 350$, bias is approximately $-7\,\%$ when only two listening circles overlap, and $+\,5\,\%$ when six overlap. In all situations, bias decreases as the number of listening circles increases. For more details, see Horrocks et al. [1].

We also examined various violations of the assumptions including nonconstant distance between listening circles, nonconstant range of detection, nonconstant density, nonuniform distribution of whales, and nonconstant probability of detection. Here, $N = 1000$, $d = 5$, $r = 10$, $\mu = 0.9$, and $\alpha = 0.001$ and there were 10,000 simulated data sets. The greatest bias in the estimation of α occurs under violation of the Poisson assumption, i.e., when individuals are not uniformly distributed across space, as occurs when density varies (scenario 6) and when individuals avoid each other (scenario 8). These results are shown in Fig. 2. For more details, see Horrocks et al. [1].

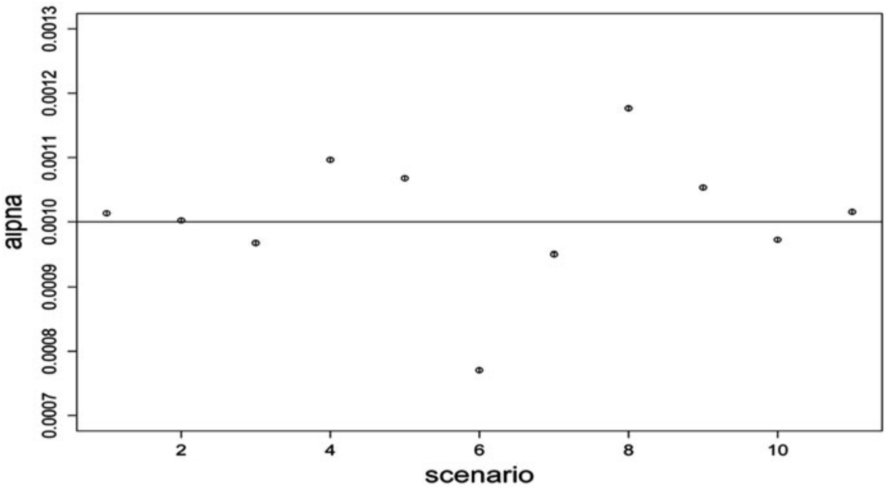


Fig. 2 Bias in α under (1) standard assumptions, (2) nonconstant distance, (3) r varies across individuals, (4) r varies across listening circles, (5) r varies across individuals and circles, (6) α varies by a factor of 2, (7) α varies by a factor of 1.5, (8) whales within a distance r avoid each other, (9) whales within $r/2$ avoid each other, (10) whales attract each other, (11) μ varies. The true value of $\alpha(0.001)$ is shown as a *horizontal line*

We now illustrate the model using data collected as part of an acoustic survey conducted in the western part of the Sargasso Sea between February 24 and March 5, 2008. The Sargasso Sea is an oceanic gyre bounded by ocean currents. It lies in the middle of the North Atlantic Ocean between 20–25° N and 30–70° W, extending westward to the Gulf Stream. As reported in [3], a hydrophone was towed behind a 12.5 m sailing vessel and was monitored approximately every half hour. Here we analyze data on sperm whales. There were 332 listening circles over approximately 1952 km, and 28 of the listening circles were positive, i.e., sperm whales were heard. Sperm whales are generally found either in groups of adult females and immatures or as single adult males. The data reported here are for groups of adult females and immatures only. Thus for these data, α will be interpreted as the number of groups of sperm whales per km^2.

We used profile likelihood to obtain 95 % confidence intervals. We obtained a maximum likelihood estimate (MLE) for α of 0.00036 groups of whales per km^2 with 95 % confidence interval (0.00016, 0.00074). The probability of detection and vocalization, μ, was estimated as 0.9 (0.64, 1) and the radius of detection, r, was estimated as 9.29 km (6.95, 15.14). An estimate of the mean group size for adult females and immature sperm whales in the Sargasso Sea is 12 individuals [4]. Given this, we estimated the density of sperm whales in this area to be 4.32 whales per 1000 km^2.

4 Conclusions

In this chapter, we have presented a method for estimating abundance of aquatic species from an acoustic binary time series. The binary data arises because it is not possible to tell how many whales are vocalizing or if the same whale is being heard more than once, and thus only imprecise data are available, namely, binary presence/absence data. In order to estimate absolute abundance from such imprecise data, it is necessary to make parametric assumptions. Due to the special design of the survey, successive observations are dependent, and this independence is exploited to allow the estimation of up to three parameters. Biological knowledge about the habits of the species, namely that they travel in groups, was used to develop an estimate of the abundance of individuals.

References

1. Horrocks, J., Whitehead, H., Hamilton, D.: A likelihood approach to estimating animal density from binary acoustic transects. Biometrics **67,** 681–690 (2011)
2. Schwarz, C.J., Seber, G.A.F.: Estimating animal abundance, review III, Stat. Sci. **14,** 427–456 (1999)
3. Whitehead, H.: Estimating abundance from one-dimensional passive acoustic surveys. The J. Wildl. Manage. **73,** 1000–1009 (2009)
4. Wong, S.N.P.: A pelagic paradox: the ecology of a top predator in an oceanic desert. Unpublished PhD thesis, Dalhousie University. (2012)

Design, Fabrication, and Testing of Hybrid Energy Harvesting Units

Mohammed Ibrahim and Armaghan Salehian

Abstract The increasing usage of mobile electrical units demands higher energy efficiencies for these devices. Self-sustaining units that harvest various forms of ambient energy can help significantly with their regular battery replacements. In this chapter two hybrid energy harvesting units are proposed that employ piezoelectric, magnetostrictive, and electromagnetic technologies to capture ambient vibrational energy. The first harvester is made of piezoelectric and magnetostrictive materials while the second harvester is composed of a piezoelectric layer and a magnet. Both proposed harvesters employ a spiral piezoelectric layer in order to reduce the compliance of the piezoelectric unit. The advantages of the first design is that it allows for more efficient harvesting over a wider range of frequencies than traditional harvesting units while the second design reduces the natural frequency of the system that results in better energy harvesting at low frequencies.

1 Introduction

Energy harvesting is becoming more important as energy sources become increasingly scarce and expensive. With recent advancements in electronic technology, sensors require less power to operate, thus ambient energy harvesting methods become potential solution for powering sensors.

This chapter focuses on harvesting ambient vibrations. Piezoelectric [1, 2], electromagnetic [3, 4], electrostatic [5, 6], and magnetostrictive [7, 8] technologies are commonly used for harvesting vibration energy.

Karimi et al. [9] designed an analytical model for vibrations analysis of spiral beams. It was concluded that the movement of the beam due to vibration is primarily torsional. Hu et al. [10] made a spiral shaped piezoelectric harvester that is actuated

A. Salehian (✉) · M. Ibrahim
Mechanical and Mechatronics Engineering Department,
University of Waterloo, Waterloo, ON, Canada
e-mail: salehian@uwaterloo.ca

M. Ibrahim
e-mail: m6ibrahi@uwaterloo.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_35

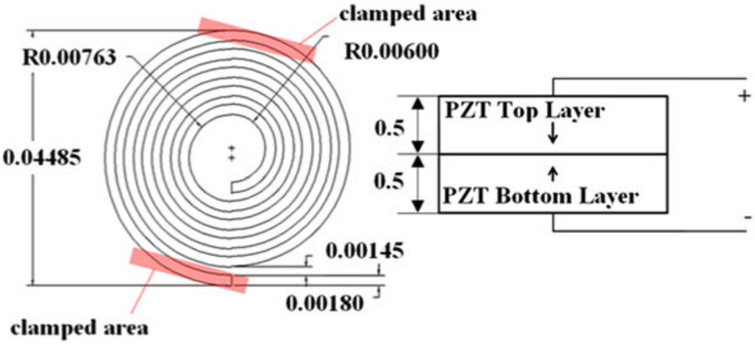


Fig. 1 Schematic of the spiral piezoelectric bimorph unit and polarization direction. Dimensions are in meter

by a 1 mN force and is fixed at the center. After optimization the resulting harvester had its first natural frequency at 50 Hz, which is relatively high for capturing ambient vibrations energy. Wang and Yuan [11] fabricated a cantilever beam with eight magnetostrictive laminates to harvest vibrations energy through Faraday's law of induction via a pickup coil. It was found that magnetostrictive material has the capability to compete with piezoelectric material for energy harvesting. Wischke et al. [12] added a magnet at the tip of a piezoelectric cantilever beam to further reduce its natural frequency while actively contribute to the harvesting.

In this chapter, consideration is given to a spiral design as well different harvesting technologies to achieve a hybrid unit with improved harvesting capabilities. The piezoelectric–magnetostrictive (PMSM) harvester uses both piezoelectric and magnetostrictive materials to achieve a wide band harvester. The piezoelectric–electromagnetic (PMAG) harvester uses piezoelectric material with a central magnet.

2 Spiral Piezoelectric Design

A schematic of the dimensions of the spiral used in this work is shown in Fig. 1. The spiral is made of two sets of half circles that have an offset center. The piezoelectric spiral is in a bimorph configuration. During vibrations, one layer is in tension while the other is in compression. Fig. 1 depicts the two layers of the piezoelectric bimorph in a series connection and the arrows indicate the directions of polarization.

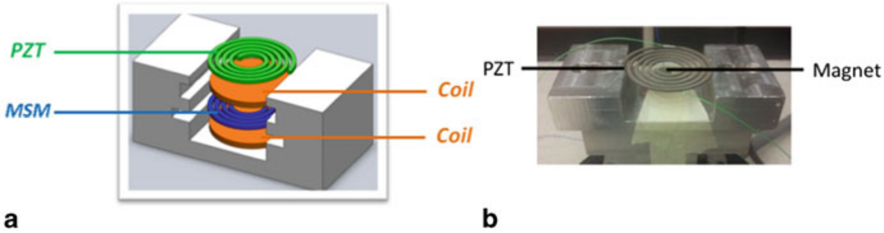


Fig. 2 **a** Schematic of the piezoelectric-magnetostrictive (PMSM) harvester and the coils. **b** PMAG harvester

3 Experiment

3.1 *Piezoelectric–Magnetostrictive Harvester*

The PMSM is made of two separate piezoelectric and magnetostrictive layers as shown in Fig. 2a. The magnetostrictive material, magnetic alloy 2605SA1 from Metglas Incorporated was used. The piezoelectric material, PSI-5A4E, was purchased from Piezo Systems.

The magnetostrictive material is fabricated using 25 μm laminate sheets. Through experimentation it was observed that combining multiple layers strengthens the magnetic field of the material; therefore 100 layers were used in order to produce larger voltage output. The MSM layers were cut into spiral shape using electric discharge machining (EDM) in the same geometries as the piezoelectric material except for the thickness. The magnetostrictive spiral was then annealed in an oven under a magnetic field in order to align the poles of the 100 layers. The layers were subsequently epoxied together. When subjected to vibrations, the magnetostrictive material produces a variable magnetic field that may be harvested in the form of electricity using a coil. A 3000 turn copper coil was employed for this purpose.

3.2 *Piezoelectric–Electromagnetic Harvester*

The PMAG harvester is composed of the piezoelectric unit discussed in Section 2 and a magnet attached to the center of the unit as shown in Fig. 2b. The Central magnet helps with reduction at the fundamental natural frequency while acting as an active harvesting unit.

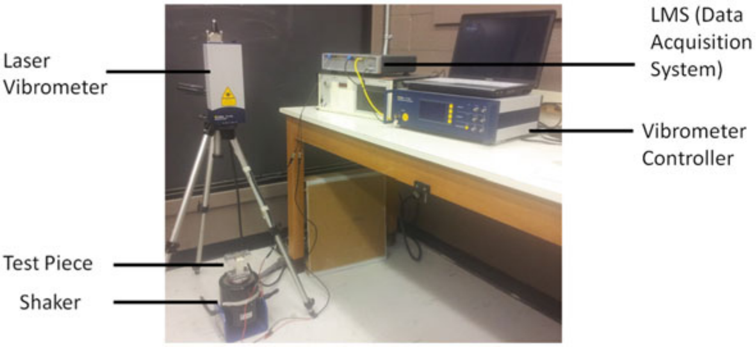


Fig. 3 Testing equipment

3.3 Testing Equipment

The test setup used for both harvesters can be seen in Fig. 3. The harvesting units were mounted on the shaker and a constant 0.3 g acceleration was held through a frequency sweep of 10–100 Hz. The data was measured through the LMS SCADA MOBILE V data acquisition system (DAS) and recorded through the Sine control module of the Test Lab software.

4 Results and Discussion

4.1 PMSM

The experimental power output for the PMSM harvester is presented in Fig. 4. The piezoelectric material has its first resonance at 24 Hz with a power output of about 6 μW. The magnetostrictive material has its first resonance at 17 Hz with a power output of 3.1 μW. The solid line in Fig. 4 indicates the total power output of the hybrid device, As shown, the frequency range for which useful energy can be harvested is wider in comparison to each material separately. This wider frequency bandwidth along with use of 2 materials for active harvesting can aid in better designs to improve harvesting capability.

4.2 PMAG

The experimental results of the PMAG harvester are presented in Fig. 5. The harvester shows a fundamental resonant frequency of 21 Hz. The piezoelectric material has a power output of about 6 μW and the magnet has a power output of 10 μW. The

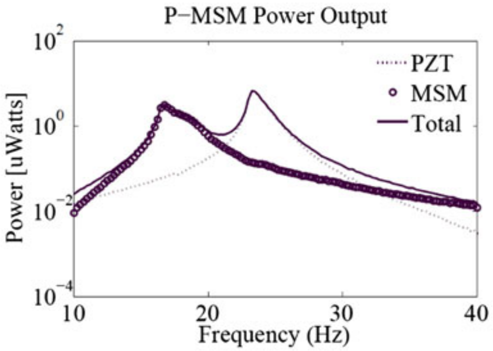


Fig. 4 Power output for PMSM harvester

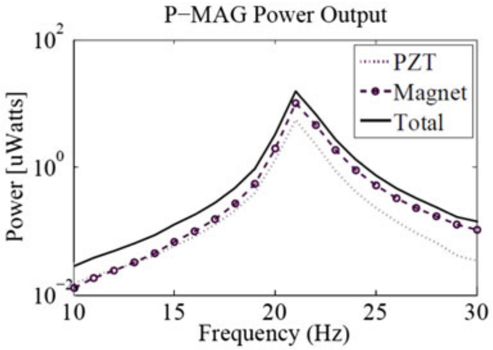


Fig. 5 Power output for PMAG harvester

experimental results indicate that even though the PMAG harvester is capable of harvesting more it has a smaller frequency bandwidth when compared to the PMSM.

5 Conclusion

Two hybrid energy harvester units are developed. One, PMSM that, operates over a wider range of frequencies compared to the piezoelectric spiral unit. The other, PMAG, has smaller natural frequency than the piezoelectric spiral while using both magnet and the spiral piezoelectric for harvesting. smaller natural frequencies are always advantages for harvesting from ambient due to small frequencies available in ambient vibrations. Experimental results indicate that combining two energy harvesting technologies results in more practical devices for harvesting ambient vibrations with higher power density.

Future work involves optimization of the spiral geometry for piezoelectric and magnetostrictive units to increase the power output.

Acknowledgments The authors are grateful for support of NSERC funding agency NSERC-DG 371472-2009. They also sincerely thank Dr. Raafat Mansour from the Electrical and Computer

Engineering department at the University of Waterloo for allowing access to his lab facilities. In addition, they would like to thank Saman Nazari for his help with annealing the magnetostrictive material.

References

1. Gilbert, J.M., Balouchi, F.: Comparison of energy harvesting systems for wireless sensor networks. Int. J. Autom. Comput. **05**(4), 334–347 (2008)
2. Roundy, S., Wright, P.K., Rabaey, J.: A study of low level vibrations as a power source for wireless sensor nodes. Comput. Commun. **26,** 1131–1144 (2003)
3. Williams, C.B., Shearwood, C., Harradine, M.A., et al.: Development of an electromagnetic microgenerator. IEE Proc. Circuits Devices Syst. **148**(6), 337–342 (2001)
4. Khaligh, A., Zeng, P., Zheng, C.: Kinetic energy harvesting using piezoelectric and electromagnetic technologies, state of the art. IEEE Trans. Ind. Electron. **57**(3), 850–860 (2010)
5. Mahmoud, M.A.E., Abdel-Rahman, E.M., El-Saadany, E.F., et al.: Electromechanical coupling in electrostatic micro-power generators. Smart Mater. Struct. **19**(2), 1–8 (2010)
6. Roundy, S., Wright, P.K., Pister, K.: Micro-electrostatic vibration-to-electricity converters. IMECE2002, November 17–22, 2002, New Orleans, 39309. ASME. (2002)
7. Wu, G., Zhang, R., Li, X., et al.: Resonance magnetoelectric effects in disk-ring (piezoelectric-magnetostrictive) composite structure. J. Appl. Phys. **110**(12), 124103 (2011)
8. Li, L., Lin, Y.Q., Chen, X.M.: CoFe2O4/Pb(Zr052Ti0.48)O3 disk-ring magnetoelectric composite structures. J. Appl. Phys. **102**(6), 064103 (2007)
9. Karimi, M.A., Yardimoglu, B., Inman, D.: Coupled out of plane vibrations of spiral beams for micro-scale applications. J. Sound Vib. **329**(26), 5584–5599 (2010)
10. Hu, H., Xue, H., Hu, Y.: A spiral-shaped harvester with an improved harvesting element and an adaptive storage circuit. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**(6), 1177–1187 (2007)
11. Wang, L., Yuan, F.G.: Vibration energy harvesting by magnetostrictive material. Smart Mater. Struct. **17**(4), 1–14 (2008)
12. Wischke, M., Masur, M., Goldschmidtboeing, F., et al.: Electromagnetic vibration harvester with piezoelectrically tunable resonance frequency. J. Micromech. Microeng. **20**(3), 1–7 (2010)

Markov Chain Monte Carlo Analysis of Trophic Cascade in Yellowstone after Reintroduction of Wolves

Darryl Johnson, David J. Klinke, Qing Wang, Morgan Condon and Zhijun Wang

Abstract In this chapter, we update a mathematical model based on the Lotka-Volterra predator–prey model to describe the elk–coyote–wolf interactions after the reintroduction of wolf in Yellowstone in 1995. A Markov Chain Monte Carlo algorithm is applied to calibrate the model parameters based on data compiled since wolves were released in the park. Our model predictions match the published experimental data very well. The objective of this study is to predict the impact of wolf reintroduction into the Yellowstone National Park on elk and coyote population.

1 Introduction

Population growth models have been widely investigated due to their great potential in aiding adaptive management for conservation purposes [1–3, 11]. Influence of harvest, climate, and wolf predation on Yellowstone elk was investigated in [12]. Valey and Boyce developed a discrete predator–prey model to describe the impact of the reintroduction of wolf in Yellowstone in 1995 on the elk population [11]. Berge and Case [3] tested the hypothesis that interference competition with wolves

Q. Wang (✉) · D. Johnson · M. Condon · Z. Wang
Department of Computer Sciences, Mathematics, and Engineering,
Shepherd University, Shepherdstown, WV 25433, USA
e-mail: qwang@shepherd.edu

D. Johnson
e-mail: nisshoku561@gmail.com

M. Condon
e-mail: void.presence@gmail.com

Z. Wang
e-mail: zwang@shepherd.edu

D. J. Klinke
Department of Chemical Engineering and Mary Babb Randolph Cancer Center,
Department of Microbiology, Immunology and Cell Biology,
West Virginia University, Morgantown, WV 25606, USA
e-mail: david.klinke@mail.wvu.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_36

limits the distribution and abundance of coyotes, and the extirpation of wolves is often invoked to explain the expansion in coyote range throughout much of North America. Predation, as reported by Forrester and Wittmer [5], was the primary proximate cause of mortality for all age classes, and was an important source of summer fawn mortality and of mortality in multi-prey, multi-predator systems. While coyote is also a major predator of calf elks, its population is greatly impacted by the reintroduction of wolf [4, 10]; there have been very few models that investigate all three species (wolf, elk, and coyote). In this study, we updated the Lotka–Volterra predator–prey model to describe the interactions between wolf, coyote, and elk in Yellowstone National Park after the release of wolves in 1995.

On the other hand, a typical ordinary differential equation (ODE) model for a biological process or phenomenon often contains dozens of parameters to be fitted against experimental data. Isolating a single parameter and attempting to ascertain its likely value is a difficult and often impossible task in a complicated system, reducing the efficacy of ODE models in describing the relationship among the postulated elements of the system from the observed data. Markov Chain Monte Carlo (MCMC) algorithms is a wide class of methods seeking to sample a probability distribution that corresponds to the distribution of likely parameter values, given the observed data and the model using a random walk in parameter space. Inclusion of a new point in the walk is conditioned on how well the corresponding model predictions match the observed data. Here, we use a variant of a widely used MCMC method, the Metropolis–Hastings algorithm, and apply it to calibrate our updated predator–prey model using published population data.

The rest of the chapter is organized as follows. Section 2 presents the modified predator–prey model with a set of assumptions incorporated into the traditional Lotka–Volterra predator–prey model. In Sect. 3, we briefly state the MCMC algorithms and the Gelman–Rubin diagnostic to determine the convergence of Markov chains. Section 4 presents numerical simulations of the proposed model with parameters calibrated against published elk–coyote–wolf population data using MCMC techniques. In Sect. 3, we discuss the conclusions of our findings based on the posterior distributions in the model parameters and predictions.

2 The Model

An ecological example involving the MCMC method is the examination of the trophic cascade that followed the reintroduction of wolves to the Yellowstone National Park in 1995. We use a model similar to the Lotka–Volterra model but with three species. In the equations below, E is the number of elks in thousands, C is the number of coyotes, W is the number of wolves, and t represents time in years.

A revised version of the Lotka–Volterra predator–prey model is described as follows:

$$\frac{dE}{dt} = k_{p1}E - p_{ec}EC - p_{ew}EW \tag{1}$$

$$\frac{dC}{dt} = -k_{d1}C + k_{p2}EC - k_{d2}CW \tag{2}$$

$$\frac{dW}{dt} = -k_{d3}W + k_{p3}EW, \tag{3}$$

where k_{p1} is the rate of growth of the elk, p_{ec} is the predation rate of coyotes on elk, p_{ew} is the predation rate of wolves on elk, k_{d1} is the natural decay rate of coyotes given lack of food, k_{p2} is the proliferation rate of coyotes, k_{d2} represents hostile incidents between wolves and coyotes that have occurred since the wolves reintroduction, k_{d3} is the natural decay rate of wolves given lack of food, and k_{p3} is the proliferation rate of the wolves given sustenance. The model assumes that

- The elks have an unlimited food supply and follow an exponential growth pattern in the absence of predators.
- The coyotes and wolves undergo exponential decay due to either natural death or emigration in the absence of prey.
- The coyotes get killed by wolves in the competition in elk-predation.
- Population sizes of coyotes and wolves increase due to elk-predation.

3 The MCMC Algorithms

MCMC algorithms are a wide class of methods seeking to sample a probability distribution from a Markov chain whose equilibrium is the probability distribution we are seeking. We use a variant of a widely used MCMC method [6, 7], the Metropolis–Hastings algorithm, and apply it to a revised predator–prey model.

The algorithm is briefly described by

1. Choose a point $x_j,\ j = 0$ in parameter space, preferably close to a realistic value.
2. Calculate the likelihood $P(Y|M(x_0))$, that you would observe experimental data similar to the simulated data based on these parameters.
3. Take a random step x_{prop} from this point that is distributed according to the proposal function, $f(X)$.
4. Calculate the probability $P(Y|M(x_{prop}))$ that the experimental data can be simulated with these new parameters.
5. Calculate acceptance probability as the ratio of the proposed step and the current step: $h = \frac{P(Y|M(x_{prop}))}{P(Y|M(x_j))}$.
6. Accept the new step with probability $min(h, 1)$. If accepted $x_{j+1} = x_{prop}$, If not accepted $x_{j+1} = x_j$.
7. Go to 3.

The proposal function, $f(X)$, can incorporate prior information about parameter ranges and correlation among the parameters. Alternatively, the proposal function can be estimated empirically from the correlation among the parameters obtained from the cumulative Markov chain [7]. Subsequent steps can be proposed that reflect this parameter structure. If the adaptation of the proposal distribution diminishes in the limit of a long chain, the distribution obtained by the Markov chain converges toward the target distribution, the posterior probability in the model parameters $P(X)$. By focusing on parameter combinations that provide predictions consistent with the observed data, convergence of the Markov chain to the target distribution can be achieved more rapidly. If the proposal function is symmetric and re-centered at each proposed step, the Markov chain is considered reversible and the acceptance probability is defined by the likelihood ratio, as shown above. Here, we used a uniform prior, meaning that parameters could take any value between 10^{-10} and 10^{10}, and a symmetric proposal distribution, that is a square gaussian matrix with the same values for the variance along the diagonal and the off-diagonal elements were set to zero. The proposal distribution was multiplied by a scalar value, where the scalar value was determined such that 20 % of the proposed steps were accepted.

To test whether a partial chain has ran long enough to hold properties sufficiently similar to the equilibrium distribution we must analyze its convergence. There are many ways to analyze convergence of Markov chains the simplest being the Gelman–Rubin diagnostic, which compares several chains starting from different initial configurations and compares their variances [7]. The driving principle behind the test is that after convergence, the behavior of some part of the chain (e.g., variance, etc.) should be similar to the whole.

4 Numerical Simulations by the MCMC Algorithm

In this section, we compare our model against observed data on the population numbers of wolf, elk, and coyote in the Northern Range of Yellowstone National Park [8]. The total number of coyotes in the park is unknown; however, we have been given two observations of their population trends specifically in the Lamar Valley [4]. We use this data and the assumption that the coyotes in other regions of the park have followed similar trends to the coyotes in the Lamar Valley.

We take as the dimensions of the configuration space the unknown parameters of our model. If an observation is given by Y_j and the corresponding differential equation response given the model parameter x is $M_j(x)$, then $\pi(x)$, the likelihood to observe Y_j given $M_j(x)$, is proportional to:

$$\prod_{j=1}^{n} \left[\frac{(Y_j - M_j(x))^T (Y_j - M_j(x))}{Max(Y_j)^2}\right]^{-\frac{N_{obs,j}}{2}},$$

where $N_{obs,j}$ is the total number of observations in experiment j and n is the number of experiments [7].

Figures 1, 2, and 3 illustrate calibration results via the MCMC algorithms.

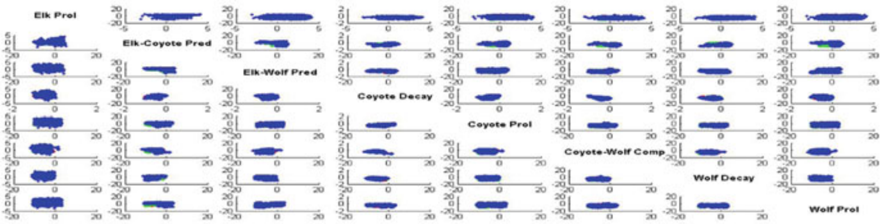


Fig. 1 This graph shows the projection of the parameter vector on each of the planes in parameter logspace. Each color (*blue*, *green*, and *red*) represents a different, independent chain. Each chain was run for 100 thousand iterations with an equal burn-in period and a thinning coefficient of 40

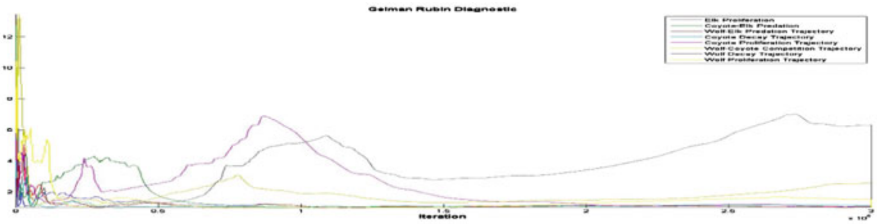


Fig. 2 The Gelman–Rubin diagnostic shows that most of the parameters in the three chains are converegent

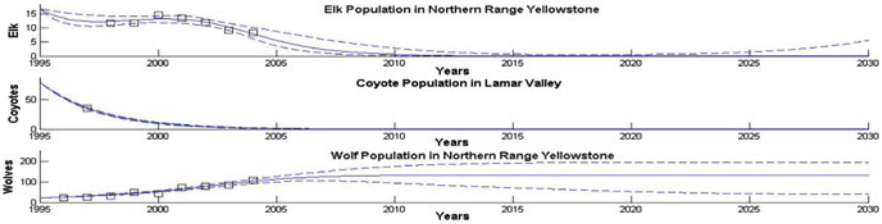


Fig. 3 Simulated results (*lines*) are compared against the experimental observations (*boxes*) used to calibrate the mathematical model. The uncertainty in the model predictions are represented by three lines: the most likely prediction is represented by the *solid lines* and the *dashed lines* represent the 95th and 5th percentile of the predicted response. The initial population are 1.67×10^4 for elk, 80 for coyote, and 21 for wolf. The parameter values of the most likely prediction are $k_{p1} = 0.467$, $p_{ec} = 7.88 \times 10^{-3}$, $p_{ew} = 7.16 \times 10^{-3}$, $k_{d1} = 0.4251$, $k_{p2} = 2.56 \times 10^{-4}$, $k_{d2} = 4.96 \times 10^{-7}$, $k_{d3} = 2.81 \times 10^{-7}$, and $k_{p3} = 0.015$

5 Discussions and Conclusions

This study updated the Lotka–Volterra predator–prey model whose parameters were calibrated against published experimental data using the MCMC techniques and Gelman–Rubin diagnostic. With the parameters derived from taking the means of three Markov chains, our model predictions match the published experimental data very well (see Fig. 3). The model suggested a sharp decrease of elk population after the release of wolves in 1995 and a slow recovery in elk population in the Northern Range Yellowstone after year 2025 with a risk of extinction. The model predicted an increase in wolf population in the Northern Range Yellowstone after the release possibly due to plenty of prey and a slowdown or decrease in number because of the decrease of prey population. Our model also indicated a rapid extinction of coyotes; this would be a result of lack of good data for coyotes. With only two data points for coyotes in the Lamar Valley found in the literature, we merely predicted a 50 % decrease over the course of 2 years in a very specific subset of the population with no available experimental data showing any recovery trend after the rapid decrease. Future work would consider the impact of human interference and animal migration.

Acknowledgments This work was supported by grants from the National Science Foundation (CAREER 1053490 to David J. Klinke II), the National Cancer Institute (R15CA132124 to David J. Klinke II), and the National Institute of General Medical Sciences of the National Institutes of Health grant as part of the West Virginia IDeA Network of Biomedical Research Excellence (P20GM103434 to QingWang).

References

1. Allen, J.J., Bekoff, M., Crabtree, R.L.: An observational study of coyote (*Canis latrans*) scent-marking and territoriality in Yellowstone National Park. Ethology **105**, 289–302 (1999)
2. Ballard, W.B., Lutz, D., Keegan, T.W., Carpenter, L.H., deVos, Jr. J.C.: Deer-predator relationships: a review of recent North American studies with emphasis on mule and black-tailed deer. Wild. Soc. Bull. **29**(1), 99–115 (2001)
3. Berger, K.M., Gese, E.M.: Does interference competition with wolves limit the distribution and abundance of coyotes? J. Animal Ecol. **76**, 1075–1085 (2007)
4. Douglas W. S., Peterson, R.O., Houston, D.B.: Yellowstone after wolves. BioScience **53**, 330–340 (2003)
5. Forrester, T.D., Wittmer, H.U.: A review of the population dynamics of mule deer and black-tailed deer *Odocoileus hemionus* in North America. Mammal Rev. **43**, 293–308 (2013). doi:10.1111/mam.12002
6. Gaimerman, D., Lopes, H. F.: Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. CRC Press, Boca Raton (2006)
7. Klinke, D.J.: An empirical Bayesian approach for model-based inference of cellular signaling networks. BMC Bioinform. **10**(371), 18 pp. (2009)
8. Nyberg, N.: Studying population trends of the grey wolf and the elk in Yellowstone National Park: discrete dynamical approach. B McQuarrie. morris.umn.edu. http://www.morris.umn.edu/academic/math/Ma4901/Sp2011/Final/NicoleNyberg-final.pdf. (2011)
9. Ripple, W.J., Wirsing, A.J., Wilmers, C.C., Letnic, M.: Widespread mesopredator effects after wolf extirpation. Biol. Conserv. **160**, 70–79. (2013)
10. Robinson, W.B.: Widespread mesopredator effects after wolf extirpation. J. Mammalogy **33**, 470–476. (1952)
11. Varley N., Boyce, M. S.: Adaptive management for reintroductions: updating a wolf recovery model for Yellowstone National Park. Ecol. Model. **193**, 315–339. (2006)
12. Vucetich, J.A., Douglas W.S., Daniel R.S.: Influence of harvest, climate and wolf predation on Yellowstone elk, 1961–2004. OIKOS. **111**, 259–270. (2005)

Discovering Forward Invariant Sets for Nonlinear Dynamical Systems

James Kapinski and Jyotirmoy Deshmukh

Abstract We describe a numerical technique for discovering forward invariant sets for discrete-time nonlinear dynamical systems. Given a region of interest in the state space, our technique uses simulation traces originating at states within this region to construct candidate Lyapunov functions, which are in turn used to obtain candidate forward invariant sets. To vet a candidate invariant set, our technique samples a finite number of states from the set and tests them. We derive sufficient conditions on the sample density that formally guarantee that the candidate invariant set is indeed forward invariant. Finally, we present a numerical example illustrating the efficacy of the technique.

1 Introduction

Model-based design (MBD) is a mathematical and visual process for designing, implementing, and testing embedded software designs for real-time control systems. MBD is rapidly becoming the pervasive design paradigm in many sectors such as automotive and avionics, but the problem of checking correctness of such designs is a highly challenging task. Of particular interest is the problem of ensuring that the system satisfies *safety* constraints, which are usually associated with a region of the state space. Analysis techniques from dynamical systems theory can be applied to such designs to verify system properties, such as those for checking stability or estimating performance bounds (see, e.g., Chap. 4 of [3]); however, these are rarely used in any but the earliest stages of the MBD process.

It is well known that a sublevel set of a Lyapunov function is a *forward invariant set*. The existence of a forward invariant set that properly contains the set of initial states, while excluding the unsafe region proves that the system is safe for

J. Kapinski (✉) · J. Deshmukh
Toyota Technical Center, Powertrain Control (Model-based Development),
1630 W. 186th St., Gardena, CA, USA
e-mail: jim.kapinski@tema.toyota.com

J. Deshmukh
e-mail: jyotirmoy.deshmukh@tema.toyota.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_37

all time. Thus, it is clear that identifying such invariant sets helps us to address the safety verification problem. A significant obstacle to this approach is that Lyapunov functions of arbitrary (nonlinear or hybrid) systems are notoriously hard to discover. Further, industrial models are often in formats lacking an analytic representation of the dynamics.

We now give a brief overview of our technique. We use an iterative procedure to construct candidate Lyapunov functions using simulation traces. The candidate Lyapunov functions are restricted to the class of polynomial functions, similar to the sum of squares (SoS) techniques described in [5]. This restriction allows us to compute a candidate forward invariant set by solving a linear program (LP). We then verify the validity of the candidate invariant set by testing over a finite number of system states. Note that, alternatively, if an analytic representation of the dynamics is given, one could verify the validity of the candidate invariant set using arithmetic solvers, as we describe in [2].

Our work builds largely on [6], where forward orbits (which are often called *simulations* or *simulation traces*) are used to seed a procedure to estimate the region of attraction (ROA) for a dynamical system. We provide the following extensions to that work: (a) we provide a procedure that uses a global optimizer to iteratively improve the quality of the candidate Lyapunov functions (by seeking initial conditions that falsify each intermediate candidate) and (b) Our technique is not restricted to the class of systems with polynomial dynamics.

2 Problem Statement

We consider autonomous nonlinear discrete-time dynamical systems of the form:

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k). \tag{1}$$

Here $\mathbf{x}$ represents state variables that take values in $\mathbb{R}^n$ and f is a nonlinear, locally Lipschitz-continuous vector field. We call $\hat{\mathbf{x}}$ the *successor* of $\mathbf{x}$ if $\hat{\mathbf{x}} = f(\mathbf{x})$. We assume that the system has a stable equilibrium point, which is, without loss of generality, at the origin. We address the following problem. Given the dynamical system (1), and a closed and bounded domain of interest $\mathcal{D} \subseteq \mathbb{R}^n$, identify a forward invariant set $\mathcal{S} \subseteq \mathcal{D}$ such that for all $\mathbf{x} \in \mathcal{S}$, $f(\mathbf{x}) \in \mathcal{S}$. We present a procedure that can identify such a set, without explicit knowledge of the vector field $f(\cdot)$. The following section describes the procedure.

3 Algorithm for Computing Invariant Sets

The procedure consists of three steps: (1) identify a candidate Lyapunov function for (1) within $\mathcal{D}$; (2) use the candidate Lyapunov function to compute a candidate invariant set; (3) certify that the candidate invariant set is a forward invariant set. We now describe each step in the process.

Identifying a Candidate Lyapunov Function Ideally, we want to discover a differentiable function v that $\forall \mathbf{x} \in \mathcal{D}$ satisfies:

$$v(\mathbf{x}) \succ 0 \tag{2}$$

$$v(\mathbf{x}) - v(\hat{\mathbf{x}}) > 0, \quad \forall \mathbf{x} \in \mathcal{D} \setminus \{0\},\ v(0) = 0. \tag{3}$$

Here, $v(\mathbf{x}) \succ 0$ means that v is positive definite, i.e., $\forall \mathbf{x} \neq 0$, $v(\mathbf{x}) > 0$, and $v(0) = 0$. The problem of identifying such a function v for the general case is of infinite dimension. We relax the problem by restricting the form of v as $v(\mathbf{x}) = \mathbf{z}^T \mathbf{P} \mathbf{z}$, where $\mathbf{z}$ is some vector of m monomials in $\mathbf{x}$ (e.g., $\mathbf{z} = [x_1\ x_1^2 x_2\ x_2^2]^T$) and $\mathbf{P} \in \mathbb{R}^m \times \mathbb{R}^m$. We use a collection of state/successor pairs to automatically produce candidate Lyapunov functions for the system. Given M pairs of points $\mathbf{x}_i, \hat{\mathbf{x}}_i$, where $i \in \{1, 2, \ldots, M\}$ and $\mathbf{x}_i \neq 0$ for all i, we formulate the following LP:

$$\begin{aligned} &\max_{\mathbf{P},\gamma}\ \gamma \\ s.t.\quad &\gamma > 0, \quad \text{and} \quad \forall i \in \{1, \ldots, M\}, \\ &v(\mathbf{x}_i) > 0 \\ &v(\mathbf{x}_i) - v(\hat{\mathbf{x}}_i) > \gamma \|\mathbf{x}_i\|^2. \end{aligned} \tag{4}$$

Any feasible solution to (4) results in a candidate Lyapunov function v that satisfies M necessary conditions for (2) and (3). We note that we could strictly enforce (2) by requiring that $\mathbf{P} \succ 0$, but this would require that a more expensive semidefinite program (SDP) be solved instead of an LP.

Once a candidate Lyapunov function is obtained from (4), we employ a *falsifier* to select state/successor pairs that can be used to improve the candidate Lyapunov function. The falsifier is a global optimizer that attempts to solve the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{x} \in \mathcal{D}}\ v(\mathbf{x}) - v(\hat{\mathbf{x}}) \\ s.t.\quad &\hat{\mathbf{x}} = f(\mathbf{x}). \end{aligned} \tag{5}$$

If the solution to (5) is less than zero, then the optimal $\mathbf{x}$ is a witness that falsifies the Lyapunov condition (3). This witness is added to the collection of state/successor pairs and (4) is solved again. This procedure continues until no falsifying witness can be found by solving (5). For our experiments, we use a *simulated annealing* algorithm to implement the falsifier. Figure 1 illustrates our iterative procedure.

We note that if both (a) the falsifier is capable of computing a global minimum and (b) the procedure in Fig. 1 halts, then the resulting candidate Lyapunov function $v(\cdot)$ is a Lyapunov function for (1). Practical falsifiers cannot reliably find a global minimum in general. Hence, we still need to verify the soundness of the forward invariant set computed using a candidate Lyapunov function obtained from this procedure, and we present a technique to do so later in this section.

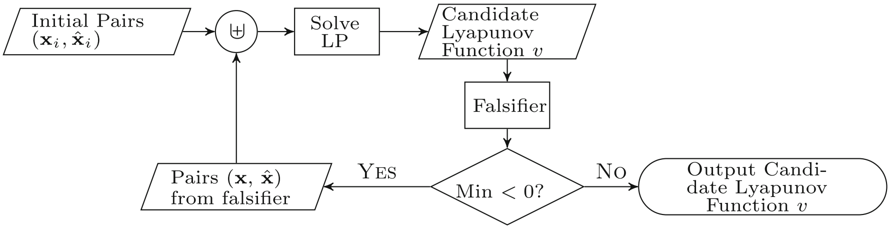

Initial Pairs
$(\mathbf{x}_i, \hat{\mathbf{x}}_i)$
$\uplus$
Solve LP
Candidate Lyapunov Function v
Falsifier
Min < 0?
YES
NO
Pairs $(\mathbf{x}, \hat{\mathbf{x}})$ from falsifier
Output Candidate Lyapunov Function v

Fig. 1 Procedure to create a candidate Lyapunov function for system (1)

Computing a Candidate Invariant Set Once we obtain a candidate Lyapunov function for (1), we can use it to obtain a forward invariant set. We formulate a convex optimization problem to maximize l such that the sublevel set $\mathcal{S} = \{\mathbf{x} | v(\mathbf{x}) \leq l\}$ is within $\mathcal{D}$. If we assume that $\mathcal{D}$ is a sublevel set of a polynomial, then standard numerical techniques can be used to obtain the optimal l, as in [1].

Verifying Soundness of the Candidate Invariant Set Below we show how to verify the soundness of the candidate invariant set computed in the previous step. The technique requires that a Lyapunov-like condition be satisfied at a finite sampling of the points in the set. First, we define a notion of sampling for a set.

Definition 1 [Delta Sampling] Given a $\delta \in \mathbb{R}_{>0}$, a δ-sampling of set $\mathcal{S} \subset \mathbb{R}^n$ is a finite set $\mathcal{S}_\delta$ such that the following holds: $\mathcal{S}_\delta \subset \mathcal{S}$; for any $\mathbf{x} \in \mathcal{S}$, there exists a $\mathbf{x}_\delta \in \mathcal{S}_\delta$ such that $\|\mathbf{x} - \mathbf{x}_\delta\| < \delta$.

The following theorem allows us to test whether a given set is forward invariant by testing a finite subset of points within the set.

Theorem 1 *[Invariant Soundness] Consider system (1), where f is locally Lipschitz with constant K_f over $\mathcal{D}$. Let $\mathcal{S} = \{\mathbf{x} | g(\mathbf{x}) \leq l\}$, where $g : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a $\mathcal{C}^1$ function that is locally Lipschitz with constant K_g over $\mathcal{S}$, and let $\mathcal{S}_\delta$ be a δ-sampling of $\mathcal{S}$. If there exists a $\gamma \in \mathbb{R}_{>0}$ such that $\delta < \frac{\gamma}{K_g \cdot K_f}$ and $\forall \mathbf{x}_\delta \in \mathcal{S}_\delta$, $g(f(\mathbf{x}_\delta)) \leq l - \gamma$, then $\mathcal{S}$ is a forward invariant set.*

Proof We prove by contradiction. Assume that $\delta < \frac{\gamma}{K_g \cdot K_f}$ and for all $\mathbf{x}_\delta \in \mathcal{S}_\delta$, $g(f(\mathbf{x}_\delta)) \leq l - \gamma$ holds, but $\mathcal{S}$ is *not* forward invariant. Then it is true that for some $\mathbf{x} \in \mathcal{S}$, $f(\mathbf{x}) \notin \mathcal{S}$. Consider the point $\mathbf{x}_\delta$ in $\mathcal{S}_\delta$ closest to $\mathbf{x}$. The Lipschitz constant for the function composition $g \circ f$ is $K_g \cdot K_f$. Applying the definition of Lipschitz continuity, we have $\|g(f(\mathbf{x})) - g(f(\mathbf{x}_\delta))\| \leq K_g \cdot K_f \cdot \|\mathbf{x} - \mathbf{x}_\delta\|$. By the definition of δ-sampling, $\|\mathbf{x} - \mathbf{x}_\delta\| < \delta$, thus we have

$$\|g(f(\mathbf{x})) - g(f(\mathbf{x}_\delta))\| < \delta \cdot K_g \cdot K_f. \tag{6}$$

Since $f(\mathbf{x}) \notin \mathcal{S}$, $g(f(\mathbf{x})) > l$, i.e., $-g(f(\mathbf{x})) < -l$. By assumption, $g(f(\mathbf{x}_\delta)) \leq l - \gamma$; adding the two inequalities, we get $g(f(\mathbf{x}_\delta)) - g(f(\mathbf{x})) < -\gamma$. By the triangle inequality, we have $\|g(f(\mathbf{x}_\delta)) - g(f(\mathbf{x}))\| > \gamma$. Combining with (6) we get:

$$\gamma < \|g(f(\mathbf{x}_\delta)) - g(f(\mathbf{x}))\| < \delta \cdot K_g \cdot K_f. \tag{7}$$

This contradicts our assumption that $\delta < \frac{\gamma}{K_g \cdot K_f}$. ■

As both δ and γ cannot be selected simultaneously, we propose an iterative procedure to determine whether the γ thus computed satisfies the condition $\delta < \frac{\gamma}{K_g \cdot K_f}$. First, a δ value is selected randomly and used to create a δ-sampling of the candidate forward invariant set $\mathcal{S}$. Next, the minimum value of $\gamma = l - v(f(\mathbf{x}_\delta))$ over the finite set $\mathcal{S}_\delta$ is computed:

$$\gamma^* = \min_{\mathbf{x}_\delta \in \mathcal{S}_\delta} l - v(f(\mathbf{x}_\delta)). \tag{8}$$

If the $\gamma^* < 0$, then the candidate $\mathcal{S}$ is not a forward invariant set (since the $\mathbf{x}_\delta$ that minimizes (8) is such that $v(f(\mathbf{x}_\delta)) > l$). If $\gamma^* > 0$ and $\delta < \frac{\gamma^*}{K \cdot K_f}$, then by Theorem 1 the candidate $\mathcal{S}$ is a forward invariant set. If $\gamma^* > 0$ but $\delta \not< \frac{\gamma^*}{K \cdot K_f}$, then we select a smaller δ such that $\delta < \frac{\gamma^*}{K \cdot K_f}$ and repeat the process.

4 Example for Computing an Invariant Set

We now present an example demonstrating the technique in Sect. 3. The following dynamical system was taken from LaSalle [4]:

$$f(\mathbf{x}) = \begin{bmatrix} \dfrac{\alpha \cdot x_2}{1 + x_1^2} \\ \dfrac{\beta \cdot x_1}{1 + x_2^2} \end{bmatrix}.$$

For this exercise, we fix $\alpha = 1.0$, $\beta = 0.9$. Fig. 2a shows the result of the procedure illustrated in Fig. 1; for the selected quadratic Lyapunov function template (i.e., $\mathbf{z} = [x_1 \;\; x_2]^T$), the procedure terminates in 5.59 s[1], giving the following candidate Lyapunov function:

$$v_{\text{LaSalle}}(\mathbf{x}) = [x_1 \; x_2] \begin{bmatrix} 368.0 & -36.0 \\ -36.0 & 396.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Next, the candidate Lyapunov function is used to candidate invariant $\mathcal{S} = v_{\text{LaSalle}}(\mathbf{x}) \leq 343.3$ (as shown in Fig. 2a); the corresponding convex program takes 2.22 s. Finally, $\mathcal{S}$ is shown to be invariant using the iterative procedure from Sect. 3. The procedure halts after two iterations (i.e., γ^* is computed twice), after 5.82 s and a cumulative total of $57,877$ sample points. Fig. 2b shows the results of this step for the example.

[1] Runtime measured on an Intel Xeon E5606 2.13 GHz Dual Processor machine, with 24 GB RAM, running Windows 7, SP1.

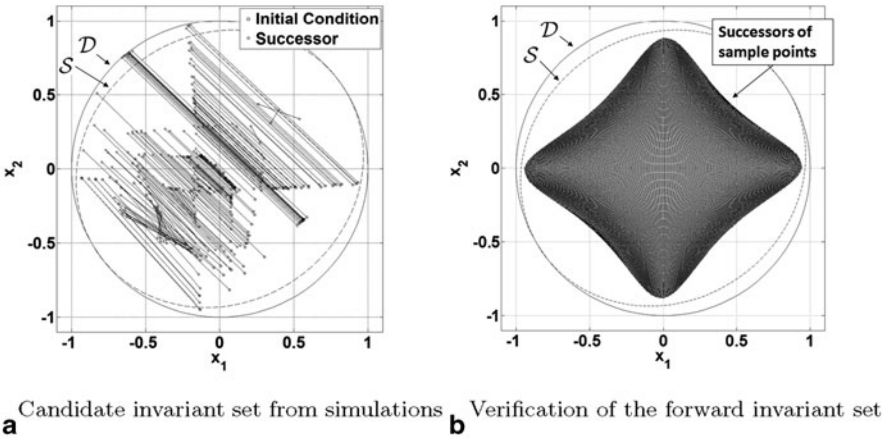


Fig. 2 LaSalle example results

5 Conclusions

We describe a numerical technique for discovering forward invariant sets for nonlinear dynamical systems using simulation traces, leveraging techniques from Lyapunov analysis, global optimization, and convex programming. The set of samples from the candidate invariant set required for verifying validity of the candidate can be prohibitively large. In future work, we will investigate satisfiability modulo theories (SMT) and interval constraint propagation solvers to symbolically test the validity of candidate invariants.

References

1. Boyd, S., Ghaoui, L.E., Feron, E., Balakrishnan, V.: Linear Matrix Inequalities in System and Control Theory, SIAM Studies in Applied Mathematics, vol. 15. SIAM (1994), Philadelphia, PA, USA
2. Kapinski, J., Deshmukh, J.V., Sankaranarayanan, S., Aréchiga, N.: Simulation-guided lyapunov analysis for hybrid dynamical systems. In: Hybrid Systems: Computation and Control (HSCC). ACM (2014), New York, NY, USA
3. Khalil, H.: Nonlinear Systems. Prentice Hall (2002), Upper Saddle River, NJ, USA
4. LaSalle, J.: The Stability and Control of Discrete Processes. Applied Mathematical Sciences. Springer (1986), New York, NY, USA
5. Parrilo, P.A.: Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization. Ph.D. thesis, California Institute of Technology (2000), Pasadena, CA, USA
6. Topcu, U., Seiler, P., Packard, A.: Local stability analysis using simulations and sum-of-squares programming. Automatica **44**, 2669–2675 (2008), Elsevier, Amsterdam, The Netherlands

Investigation of Salts Behavior at Liquid–Liquid Interfaces

N. P. Khiabani, A. Bahramian, M. Soltani, P. Pourafshary, K. Sarikhani, P. Chen and M. R. Ejtehadi

Abstract We have used molecular dynamics simulation to investigate hydrophilic–hydrophobic interfaces between calcium chloride ($CaCl_2$) aqueous solutions and normal hexane. The results demonstrate the increasing impact of salt concentration on the liquid–liquid interfacial tension, hence, negative adsorption of $CaCl_2$ according to Gibbs adsorption isotherm. Moreover, we calculated the density profiles of hexane, water, and the counter ions. The results reveal an electrical double layer near the interface and the less affinity of calcium cations toward the interface than that of chloride anions. Orientation of water molecules at the studied concentrations may result in developing a positively charged interface and, consequently, accumulation of anions close to the charged interface. Our calculations show that the interfacial width decreases by increasing salt concentration. Therefore, consistent with the calculated interfacial tension (IFT) data, aqueous salt solutions are less miscible in normal hexane at higher salt concentrations.

N. P. Khiabani (✉) · A. Bahramian
Department of Chemical Engineering, University of Tehran, Tehran 11365-4563, Iran
e-mail: nahidkhiabani@gmail.com; e-mail: nahidkhiabani@sharcnet.ca

A. Bahramian
e-mail: abahram@ut.ac.ir

N. P. Khiabani · M. Soltani · K. Sarikhani · P. Chen
Waterloo Institute for Nanotechnology, University of Waterloo, Waterloo, ON, Canada N2L 3G1

M. Soltani
Department of Mechanical Engineering, K. N. T. University of Technology, Tehran, Iran

Division of Nuclear Medicine, Johns Hopkins University, School of Medicine, Baltimore, MD, USA

P. Pourafshary
Petroleum and chemical engineering department, Sultan Qaboos University, Muscat, Oman

M. R. Ejtehadi
Departement of Physics, Sharif University of Technology, P. O. Box 11155-9161, Tehran, Iran

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_38

1 Introduction

Surface thermodynamics is one of the important branches of science because two immiscible phases of materials may come together in many processes. Many studies have dealt with the characterization of liquid–liquid interfaces because of their relevance in a lot of industrial applications such as oil industry, drug delivery, and nanoparticle synthesis. For instance, the interfacial tension of an oil–water interface is crucial to analyze the enhancement of oil recovery in petroleum reservoirs. Also, due to the existence of brine in a lot of reservoirs [1], the analysis of aqueous salts' effect on the oil–water interfacial tension is necessary. Additionally, a detailed understanding of microscopic structure of all species at these interfaces is important to optimize the relevant processes such as oil recovery methods. There is a chance to investigate molecular mechanisms as well as interpret many experimental results using molecular simulation. However, literature shows there are relatively few molecular simulation results have been established to the question of what happen at liquid-liquid interfaces in the existence of counter ions. Benjamin studied ion transfer dynamics across two immiscible liquids. The results are in reasonable agreement compared to the results of a diffusion equation solution for an ion moving in an external field [2]. Recently, Zhang and Carloni [3] looked into the effect of sodium chloride (NaCl) and potassium chloride (KCl) on oil–water interface via molecular dynamics (MD) simulation. They presented the effect of the ions on first and second layers of water at the interface, the residence time of the ions also the interfacial tension. They found: although NaCl has no effect on the interfacial structure, KCl can change it. Also, they showed that the effect of these salts on the interfacial tension and residence time of the interfacial molecules are almost the same.

In this study, we investigate the effects of calcium chloride ($CaCl_2$) as a typical bivalent salt on the interfacial behavior of water–hydrophobic interfaces. Hexane was selected as a hydrophobic liquid. We begin with an overview of our MD simulation details, followed by the estimation of the aqueous salts solutions–hexane interfacial tension and the description of density profiles.

2 Simulation Details

In this study, all MD simulations were carried out with the large-scale atomic/molecular massively parallel simulator (LAMMPS) code [4]. The total potential energy was applied as follows:

$$E_{\mathrm{total}} = E_{\mathrm{vdW}} + E_{\mathrm{Q}} + E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{torsion}} \tag{1}$$

where E_{total}, E_{vdW}, E_{Q}, E_{angle}, and E_{torsion} are the total energy, the van der Waals, electrostatic, bond-stretching, angle-bending, and torsion-energy components, respectively. The selected force field parameters are according to literature data [5–7].

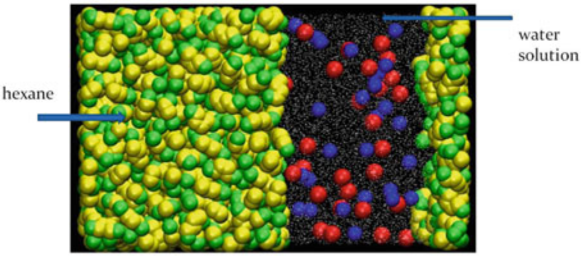


Fig. 1 A typical molecular dynamics simulation box: hexane, green and yellow particles, water, white points, salt counterions, blue and red particles

The integration of the equations of motion was performed using the Verlet algorithm [8] with a time step of 1.0 fs. The temperature was set 25°C. To control the temperature, a Nose–Hoover type thermostat [9, 10] was used. Also, the barostat type to set a 1 atm pressure was Nose–Hoover. The particle–particle particle–mesh Ewald (PPPM) method was used to calculate the long-range electrostatic interactions. The cutoff distance is considered 10 Å for both Lennard-Jones and electrostatic interactions.

3 Results and Discussion

In this section, we present both our interfacial tension results. We also discuss on the density profiles of all species in our studied systems as well as the interfacial thicknesses.

3.1 *Interfacial Tension*

At a liquid–liquid interface arises from the difference between normal and tangential components of pressure at the interface compared to the bulk for each liquid. In this study, we calculated the interfacial tension at our aqueous salt solution–hexane interface perpendicular to the z-axis using its mechanical definition [11] as follows:

$$\sigma = \frac{1}{2}\left\langle P_{zz} - \frac{1}{2}(P_{xx} + P_{yy})\right\rangle L_z \tag{2}$$

where P_{xx}, P_{yy}, and P_{zz} are the components of pressure tensor and L_z is the simulation box length in the direction perpendicular to the interface (Fig. 1).

The interfacial tension results (Fig. 2) show an increasing trend versus salt concentration increasing which is in agreement with experimental data [12]. In order to explain this phenomena, we can use the traditional Gibbs adsorption equation [13]. Using the Gibbs adsorption equation at constant temperature and the definition of chemical potential in electrochemical systems [14] we can conclude that the Gibbs adsorption isotherms would have a negative slope in case of our studied systems.

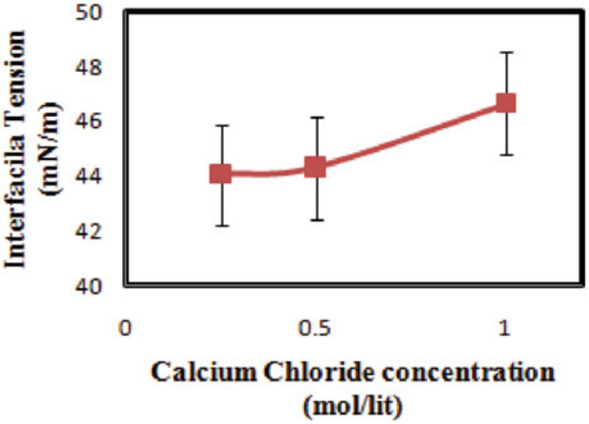


Fig. 2 Molecular dynamics simulation results of calcium chloride solution-hexane: Interfacial tension and Interfacial thickness

The fact directly evident according to the interfacial tension results (Fig. 2) and Gibbs adsorption isotherms is the negative adsorption of $CaCl_2$. The slope of interfacial tension values versus the salt concentrations is positive. So, according to the negative slope of Gibbs adsorption isotherms, $CaCl_2$ has negative adsorption at hexane–water interface and is not a surface active agent. We think that the negative adsorption of $CaCl_2$ should be related to the hrydration energy of ions. According to the continuum Born theory [15], the excess free energy of hydrated ions embedded in a liquid increases as the dielectric constant of the liquid media decreases. On the other hand, the dielectric constant of liquid solutions is lower at the interface compared to the bulk [16]. This means that ions are not interested to dissolve into the interface less interested to dissolve at the interface. Therefore, the ions with negative excess free energy of hydration such as the ions in this study [17] have negative adsorption.

3.2 Density Profiles

To obtain the density profiles of our simulated systems, we divided the simulation box into 1 Å thick slabs parallel to the xy plane and determined all the above species distribution functions in each slab. This could be a satisfied definition of the density profiles. Typically, the distribution function curve of hexane–1 molar $CaCl_2$ solution is presented in Fig. 3. It is clear that the simulated systems consist of two immiscible fluids. Another noticeable point in Fig. 3 is the existence of two peaks close to the hexane–water interfaces. The peaks are related to the counter ions net charge of the systems. They show the construction of an electrical double layer near the interface. In other words, at the studied concentrations, anions are accumulated closer to the interface than cations. Water molecules at hydrophobic interfaces are rearranged [18]. We think the asymmetric water molecules with the new arrangement may lead to the construction of an electrical double layer.

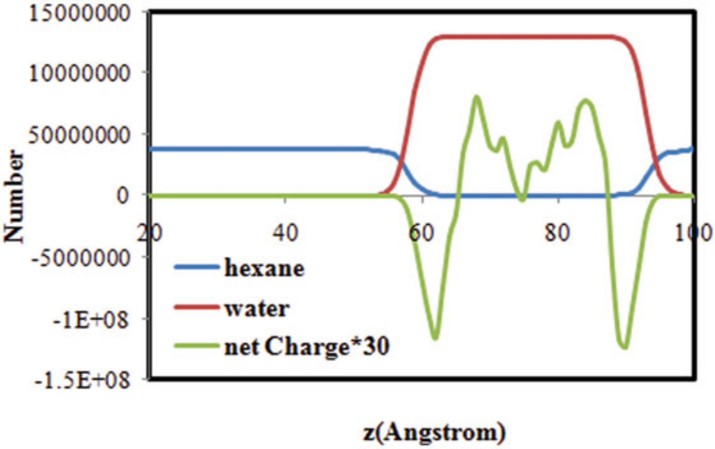


Fig. 3 Histograms of hexane, water, and the net charge of simulated hexane–1 molar $CaCl_2$ solution system

The results of the hexane–salt solution interfacial width (Fig. 2) at different salt concentrations demonstrate decreasing the interfacial thickness with increasing salt concentration. This result is consistent with the interfacial tension results and show more miscibility of hexane–salt solution at lower concentrations.

4 Conclusion

We have reported some molecular dynamics simulation results of $CaCl_2$ effect on hexane–water interface. The results at concentrations above 0.25 molar of $CaCl_2$ demonstrate the increasing of hexane–water interfacial tension. The comparison between IFT results and Gibbs adsorption equation show a negative absorption of this salt above 0.25 molar of salt concentration. The density profiles of all species at the studied interfaces show an existence of two immiscible fluids water solution and hexane, also a creation of a double layer closer to hexane–water interface. So we can conclude that the arrangements of water molecules near the interface might lead to a nonzero charge at the interface. Hence, there should be a competition between all the counter ions in these systems to be closer to the interface. The results show us that in the studied systems, anions are closer to the interface. Also, consistent with the interfacial tension results, the interfacial thickness decreases with the increasing of salt concentration.

Acknowledgement The supercomputer resources were provided by SHARCNET, Ontario, Canada. We would appreciate all SHARCNET staff members specially Dr. Pawel Pomorski for his effective comments. We also appreciate Mahmoud Ghaznavi, Mohammad Molammadi, and Parisa Sadat Mousavi for their cooperation.

References

1. Danesh, A.: Pvt and Phase Behaviour of Petroleum Reservoir Fluids. Elsevier Science, The Netherlands (1998)
2. Benjamin, I.: Dynamics of ion transfer across a liquid–liquid interface: a comparison between molecular dynamics and a diffusion model. J. Chem. Phys. **96**, 577 (1992)
3. Zhang, C., Carloni, P.: Salt effects on water/hydrophobic liquid interfaces: a molecular dynamics study. J. Phys. Condens. Matter (An Institute of Physics Journal) **24**, 124109 (2012)
4. Plimpton, S.: Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. **117**, 1–19 (1995)
5. Yan, H., Yuan, S.-L., Xu, G.-Y., Liu, C.-B.: Effect of Ca2+ and Mg2+ ions on surfactant solutions investigated by molecular dynamics simulation. Langmuir (The ACS journal of surfaces and colloids) **26**, 1044859 (2010)
6. Kalyanasundaram, V., Spearot, D.E., Malshe, A.P.: Molecular dynamics simulation of nanoconfinement induced organization of n-decane. Langmuir (The ACS Journal of Surfaces and Colloids) **25**, 755360 (2009)
7. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L.: Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. **79**, 926 (1983)
8. Swope, W.C.: A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. J. Chem. Phys. **76**, 637 (1982)
9. Hoover, W.G.: Canonical dynamics: equilibrium phase-space distributions. Phys. Rev. A **31**(331):16951697 (1985)
10. Nose, S.: A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys. **81**, 511 (1984)
11. Kirkwood, J.G., Buff, F.P.: The statistical mechanical theory of surface tension. J. Chem. Phys. **17**, 338 (1949)
12. Aveyard, R., Saleem, M.: Interfacial tensions at alkane-aqueous electrolyte interfaces. J. Chem. Soc. Faraday Trans. **1**(72):1609–1617 (1975)
13. Defay, R., Pringoging, I., Bellemans, A., Everett, D.H.: Surface Tension and Adsorption. Wiley, New York (1966)
14. Newman, J.S.: Electrochemical Systems. Wiley, New Jersey (1991)
15. Kalidas, C., Hefter, G., Marcus, Y.: Gibbs energies of transfer of cations from water to mixed aqueous organic solvents. Chem. Rev. **100**, 81952 (2000)
16. Cherepanov, D.A., Feniouk, B.A., Junge, W., Mulkidjanian, A.Y.: Low dielectric permittivity of water at the membrane interface: effect on the energy coupling mechanism in biological membranes. Biophys. J. **85**, 130716 (2003)
17. Migliore, M., Corongiu, G., Clementi, E., Lie, G.C.: Monte Carlo study of free energy of hydration for Li+, Na+, K+, F, and Cl with ab initio potentials. J. Chem. Phys. **88**, 7766 (1988)
18. Hore, D.K., Walker, D.S., Richmond, G.L.: Water at hydrophobic surfaces: when weaker is better. J. Am. Chem. Soc. **130**, 18001 (2008)

Monte Carlo Study of the Random Image Area Estimation by Pairwise Comparisons

W.W. Koczkodaj, A. Almowanes, T. Kakiashvili and G. Duncan

Abstract This study presents experimental results of gaining the accuracy of 18.4 % when the pairwise comparisons method was used instead of the direct method for area estimation of random images. Random images were produced by deblurring the Gaussian blur applied to randomly generated polygons. Participants were asked to estimate the areas of five random images by using an online questionnaire. Images have been compared to a provided unit of measure and in pairs. Our intensive Internet searches could not find another Monte Carlo experimentation for 2D case conducted in the past.

1 Introduction

Random images with smooth-looking edges were used in our Monte Carlo study. Such random images that were not too difficult to estimate their area. For it, we used a simple heuristic for generating these placated nice random images based on a modified technique in [9] posted in 2008. In reality, no one can categorically say what a nice image is. However, we can recognize nice images once we see them and more importantly, we can generate them. Smoothing the edges by deblurring help us to generate such images. However, this study is about accuracy, not the random image generation and the "quality" of randomness was not the subject of our investigation.

The pairwise comparisons is a useful method especially for processing subjective data. Its main goal is to establish the relative preference of n stimuli in situations where it is impractical to provide estimates for the stimuli [3]. The pairwise comparisons method can always be used to reach final conclusions elegantly. This method is of considerable importance in situations where direct measurements are impossible to perform. It provides a natural and a powerful tool for decision making. It is a natural approach for processing subjectivity, although objective data can also be processed this way. By common sense, and for any type of comparisons, taking two

W. W. Koczkodaj (✉) · A. Almowanes · G. Duncan
Computer Science, Laurentian University, Sudbury, ON, Canada
e-mail: wkoczkodaj@cs.laurentian.ca

T. Kakishvili
Brain Research, Baycrest, Toronto, ON, Canada

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_39

criteria or alternatives at a time works better than taking all of them at once. Evidently, handling multiple things at once is more difficult. The pairwise comparisons method is often used to subjectively compare objects. In particular, this method is used to compare objects that are difficult or impossible to measure. For example, there is no defined measure unit for the public satisfaction. The pairwise comparisons method is used for ranking all kinds of preferences and decision making. In some situations, it is the only feasible method where subjectivity is a dominant factor for a decision making.

To perform the random image Monte Carlo accuracy testing of pairwise comparisons, an online questionnaire was implemented and acted as our data collection method. Participants were asked to estimate areas of five images using a provided unit. In addition, they were asked to compare the images in pairs. The average error rate was then calculated for both and compared. The results were encouraging as the gain of accuracy reached 18.4 % when the pairwise comparisons method was used. To our own knowledge and based on an intensive search, this is the first Monte Carlo study for 2D accuracy testing of pairwise comparisons.

2 The Survey Design

Our 2D Monte Carlo experimentation for testing the pairwise comparisons method accuracy is based on using random images. The former 1D experiment in [7] was based on randomly generated bars. In [1], random images were used but of equal area. Participants related the areas of five randomly generated images of equal area. A reference unit area was also displayed along with the images. Respondents' average error when estimating the area using the unit square was 25.75 %. Nevertheless, the error went down to 5.51 % when the images were compared in pairs. It is a much better improvement percentage than the 1D case where bars were used [7]. The experiment demonstrated in [1] is the first 2D statistical experiment showing that the pairwise comparisons method improves accuracy but it was conducted for random images equal in size. In [1], a sample of 179 participated in the study. In the first part of that experiment, they were asked to estimate the area of five randomly generated images of equal areas in units. Of course, respondents were not told that the images were equal in area. The images were presented in an overhead screen and participants took, on average, 10–15 s to estimate the area of each image. In the second part, the images were shown in pairs. Ten pairs were shown and similarly it took 10–15 s to compare each pair. For each pair, participants were asked which image is larger. They also had the option to respond if they believed that a pair was equal.

Generating random images is based on deblurring in [4, 5]. In 2008, an implementation in Photoshop has been posted on the Internet [9]. A "graphical" type of a questionnaire has been designed, implemented, and programmed in Hypertext Preprocessor (PHP). The questionnaire was posted on a web page for the data collection process. The following section provides a detailed description of the data acquisition.

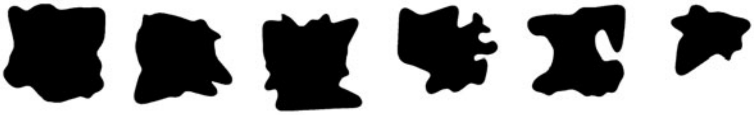


Fig. 1 Randomly generated images with unequal area sizes

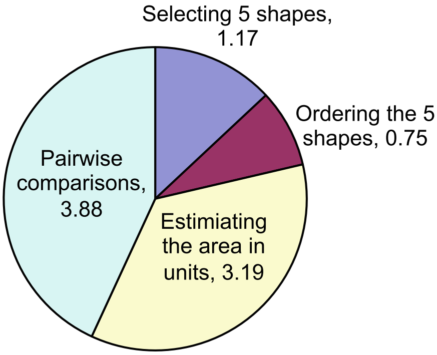
2.1 Data Acquisition Application

There are 93 recorded observations used in this experiment. There was no particular procedure for selecting participants. Only the date, time, and participants' answers were recorded. The email was also recorded only if participants asked for the results to be sent to them when the study will be completed. No Internet Protocols (IPs) or any personal identification were stored. In the first part of the experiment, participants were asked to choose 5 images from a pool of 70 images similar to the images shown in Fig. 1. They were rescaled to a smaller size (63 × 63) to make all 70 images fit the screen.

Users were asked to put in order the five randomly generated images from the largest to the smallest, where the largest gets the value of 1 and the smallest gets 5. This is to ensure that the user is able to distinguish the visible size difference among the images. In addition, it gives the ability to be consistent in the way the pair of images is displayed on the ten pairwise comparisons screens. The system allows the user to proceed to the area estimation in units page only if the ordering is correct. Otherwise, they would need to select five new images. We decided for the square unit, used in the direct method, to be of size 1600 pixels. That is a 40 × 40 unit square. The user can only input valid numeric values. If the user inputs an invalid value, an appropriate error message will be shown. If a value is valid and the submit button is clicked, the user will be taken to the next page. In the last part of the experiment, participants were shown two of the five random images side by side (pairwise comparisons). The larger image is always displayed on the left side. There were ten unique pairs that can be formed from the five images. So, ten comparisons were performed.

Polygons are then generated and filled with black and a Gaussian blur is applied to make rough edges smooth. Afterward, a threshold to transform gray pixels to black or white is used. The next step was to scale all 70 images to make them equal in area with $< 0.1\%$ margin at most. The areas are then recorded and saved to a MySQL database for easy access through PHP. We also needed to be sure that the five selected images are displayed to the user in 1–5 ratio from largest to smallest. That is why we performed the previous step of rescaling all images to approximately equal in area images and then applying a new random scale to have the five images in a 1–5 ratio. This can be done by manipulating how the image is displayed in the browser. Next, images are displayed on the ranking screen in no particular order. The user then orders them from largest to smallest.

Fig. 2 A pie chart that shows the average time taken to complete each task in minutes



2.2 Computing the Survey Results

The collected data have been transformed into a pairwise comparisons matrix M of the size 5 by 5:

$$M = \begin{bmatrix} 1 & m_{12} & \cdots & m_{1n} \\ \frac{1}{m_{12}} & 1 & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m_{1n}} & \frac{1}{m_{2n}} & \cdots & 1 \end{bmatrix}.$$

We used the theory presented in [6] as the distance-based inconsistency, extended in [2], and finally simplified in [8] as:

$$ii = 1 - \min(x * z/y, y/x/z), \tag{1}$$

for a triad (x, y, z) with all strictly positive coordinates.

The average error rate when estimating the area of random images in units (direct method), is 30.3 % for the 93 observations. On the other hand, the average error rate is only 11.96 % when the pairwise comparisons method is used, and this can be seen in Fig. 3. The gain of accuracy here is approximately 18.4 %. The results are highly encouraging. The drop of estimation error, from 30.3 to 11.96 % (see Fig. 4), is even more spectacular than the 1D case reported in [7]. It is evident that the accuracy improves when random images' area estimation using the pairwise comparisons method is enforced.

As shown in Fig. 2, the total average time that the participants needed to complete all tasks, is approximately 9 min. Although the average time taken to complete both the direct and pairwise comparisons methods are similar, the accuracy improves dramatically when the pairwise comparisons method is used.

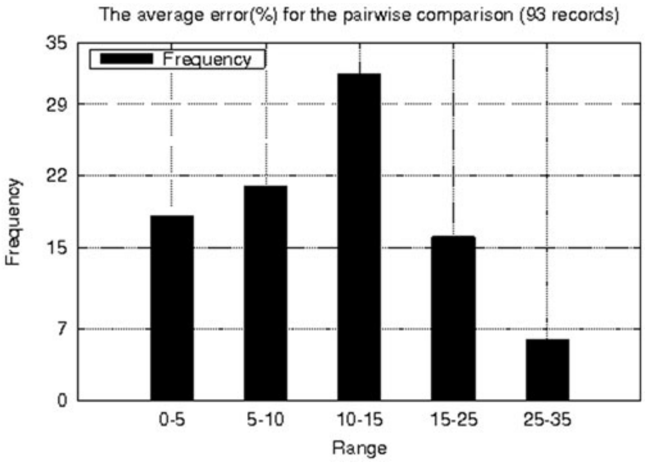
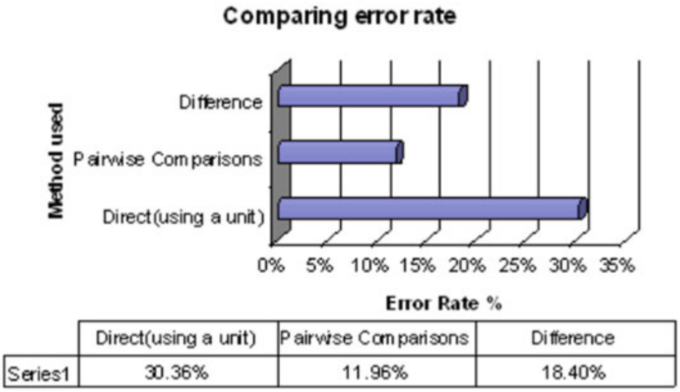


Fig. 3 Histogram showing the average error when using the pairwise comparisons method



	Direct(using a unit)	Pairwise Comparisons	Difference
Series1	30.36%	11.96%	18.40%

Fig. 4 Comparing the average error rate when using the pairwise comparisons and the direct methods for area estimation of random images

3 Conclusion

The results of our Monte Carlo experiment strongly favor the pairwise comparisons method over the direct method. The average error for the pairwise comparisons is nearly 11.96 versus 30.3 % when the direct method is used. The gain of accuracy, which is the difference between the errors derived from the direct method and the pairwise comparisons method, is around 18.4 %. It is even more impressive than the 1D case reported in [7] conducted 18 years ago. It is also worth mentioning that the average time taken to complete both the direct and pairwise comparisons methods was close, but the accuracy improves dramatically when the pairwise comparisons method is used.

Acknowledgement A. Almowanes would like to acknowledge the endeavors of the sponsor, the Kingdom of Saudi Arabia, Ministry of Higher Education.

References

1. Adamic, P., Kakiashvili, T., Koczkodaj, W.W., Babiy, V., Janicki, R., Tadeusiewicz, R.: Pairwise comparisons and visual perceptions of equal area polygons. Percept. Motor Skills **108**(1):37–42 (2009)
2. Duszak, Z., Koczkodaj, W.W.: Generalization of a new definition of consistency for pairwise comparisons. Inf. Process. Lett. **52**(5):273–276 (1994)
3. Herman, M., Koczkodaj, W.W.: A Monte Carlo study of pairwise comparison. Inf. Process. Lett. **57**(1):25–29 (1996)
4. Hummel, R.A., Kimia, B., Zucker, S.W.: Deblurring Gaussian blur. Comput. Vision Gr. Image Process. **38**(1):60–80 (1987)
5. Immerkaer, J.: Use of blur-space for deblurring and edge-preserving noise smoothing. IEEE Trans. Image Process. **10**(6):837–840 (2001)
6. Koczkodaj, W.W.: A new definition of consistency of pairwise comparisons. Math. Comput. Model. **18**(7):79–84 (1994)
7. Koczkodaj, W.W.: Statistically accurate evidence of improved error rate by pairwise comparisons. Percept. Motor Skills (1996). doi:10.2466/pms.1996.82.1.43
8. Koczkodaj, W.W., Szwarc, R.: On axiomatization of inconsistency indicators in pairwise comparisons. CoRR abs/1307.6272 (2013)
9. RandomMetropolis: Tutorial—smooth edges in Photoshop! Available via YouTube. http://www.youtube.com/watch?v=m9hicX0gyXI (2008) Accessed 1 Feb 2013

Controllability of Second Order Impulsive Differential Systems in Banach Spaces

Meili Li and Jungang Tian

Abstract This chapter shows the controllability of second order impulsive differential systems in Banach spaces. Sufficient conditions for the controllability are obtained by using the theory of strong continuous cosine families and the contraction mapping principle. Particularly, the compactness of the cosine family of operators is not needed in this chapter.

1 Introduction

In this chapter, we study the controllability of second order impulsive differential systems of the form

$$\begin{aligned} &x''(t) = Ax(t) + Bu(t) + f(t, x(t), x'(t)), \;\; t \in J, \;\; t \neq t_k, \\ &\Delta x|_{t=t_k} = I_k^1(x(t_k)), \;\; k = 1, \cdots, m, \\ &\Delta x'|_{t=t_k} = I_k^2(x'(t_k^+)), \;\; k = 1, \cdots, m, \\ &x(0) = x_0, \; x'(0) = y_0, \end{aligned} \tag{1}$$

where $J = [0, b]$, the state $x(\cdot)$ takes values in Banach space X with the norm $\|\cdot\|$, $u(\cdot) \in L^2(J, U)$ is the control function, U is a Banach space, A is the infinitesimal generator of a strongly continuous cosine family $\{C(t) : t \in R\}$ on X, $B : U \to X$ is a bounded linear operator. The functions $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = b$, $\Delta x|_{t=t_k} = x(t_k^+) - x(t_k^-)$, $\Delta x'|_{t=t_k} = x'(t_k^+) - x'(t_k^-)$, and $I_k^j : X \to X$, $j = 1, 2$, $f : J \times X \times X \to X$ are appropriate continuous functions to be specified later.

The theory of impulsive differential systems has become an important area of investigation in recent years, stimulated by their numerous applications to problems from mechanics, electrical engineering, medicine, biology, ecology, etc. (see [1, 6] and references therein).

M. Li (✉) · J. Tian
Department of Applied Mathematics, Donghua University, Shanghai, P. R. China
e-mail: stylml@dhu.edu.cn

J. Tian
e-mail: tianjg530@163.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_40

The problem of controllability of second order differential systems and inclusions has been studied by many researchers [2–5, 7, 8]. Kang et al. [5] studied the controllability of second order differential inclusion in Banach spaces. With the help of a fixed point theorem for condensing maps due to Martelli, the authors considered the damped term $x'(\cdot)$ and found a control $u(\cdot)$ in $L^2(J,U)$ such that the solution satisfies $x(b) = x_1$ and $x'(b) = y_1$. Chang et al. [3] investigated the controllability of second order differential and integrodifferential inclusions without assuming the compactness condition on the cosine family $\{C(t) : t \in R\}$ and sine family $\{S(t) : t \in R\}$. Balachandran et al. [2] pointed out an error in the paper [5]. They derived an additional condition for the controllability of second order differential inclusion in Banach spaces. Sakthivel et al. [8] studied the controllability of second order nonlinear impulsive differential systems using a fixed point analysis approach. Motivated by Balachandran et al. [2], we study the controllability for the second order impulsive differential systems in a Banach space. The method similar to that of Balachandran et al. [2]. Thus, our results extend those of Balachandran et al. [2], Chang et al. [5] and Sakthivel et al. [8].

2 Preliminaries

Definition 1 (see [9, 10]). A one parameter family $\{C(t) : t \in R\}$, of bounded linear operators in the Banach space X is called a strongly continuous cosine family iff (i) $C(s+t)+C(s-t) = 2C(s)C(t)$ for all $s,t \in R$; (ii) $C(0) = I$; (iii) $C(t)x$ is strongly continuous in t on R for each fixed $x \in X$.

The strongly continuous sine family $\{S(t) : t \in R\}$, associated with $\{C(t) : t \in R\}$, is defined by $S(t)x = \int_0^t C(s)x\mathrm{d}s$, $x \in X$, $t \in R$. Throughout this chapter, A is the infinitesimal generator of a strongly continuous cosine family $\{C(t) : t \in R\}$, of bounded linear operators from X into itself. Moreover, M and N are positive constants such that $\|C(t)\| \le M$ and $\|S(t)\| \le N$ for every $t \in J$. $[D(A)]$ is the space $D(A) = \{x \in X : C(t)x$ is twice continuously differentiable in $t\}$, endowed with the norm $\|x\|_A = \|x\| + \|Ax\|$, $x \in D(A)$. Define $E = \{x \in X : C(t)x$ is once continuously differentiable in $t\}$. E endowed with the norm $\|x\|_E = \|x\| + \sup_{0\le t\le 1} \|AS(t)x\|$, $x \in E$, is a Banach space. The operator-valued function $\mathcal{H}(t) = \begin{bmatrix} C(t) & S(t) \\ AS(t) & C(t) \end{bmatrix}$ is a strongly continuous group of bounded linear operators on the space $E \times X$ generated by the operator $\mathcal{A} = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}$ defined on $D(A) \times E$. From this, it follows that $AS(t) : E \to X$ is a bounded linear operator and that $AS(t)x \to 0$ as $t \to 0$ for each $x \in E$.

Define $PC(J,X) = \{u : J \to X,\ u(t)$ is continuous for $t \in J,\ t \ne t_k,$ and $u(t_k^+), u(t_k^-)$ exist and $u(t_k) = u(t_k^-),\ k = 1,\cdots,m\}$.

$PC^1(J,X) = \{u \in PC(J,X),\ u(t)$ is continuous differential for $t \in J,\ t \ne t_k$, and $u'(t_k^+), u'(t_k^-)$ exist and $u'(t_k) = u'(t_k^-),\ k = 1,\cdots,m\}$.

Obviously, $PC(J,X)$ is a Banach space with the norm $\|u\|_{PC} = \sup_{t\in J}\|u(t)\|$, and $PC^1(J,X)$ is also a Banach space with the norm $\|u\|_{PC^1} = \max\{\|u\|_{PC}, \|u'\|_{PC}\}$, where $\|\cdot\|$ is any norm of X.

Definition 2 A function $x \in PC^1(J,X)$ is said to be a mild solution of (1) if the impulsive conditions in (1) are satisfied and

$$\begin{aligned} x(t) = {} & C(t)x_0 + S(t)y_0 + \int_0^t S(t-s)[Bu(s) + f(s,x(s),x'(s))]\mathrm{d}s \\ & + \sum_{0<t_k<t} C(t-t_k)I_k^1(x(t_k)) + \sum_{0<t_k<t} S(t-t_k)I_k^2(x'(t_k^+)),\ \ t \in J. \end{aligned} \tag{2}$$

Definition 3 System (1) is said to be exactly controllable on the interval J, if for every $x_0 \in E, y_0 \in X$ and $x_1, y_1 \in X$, there exists a control $u \in L^2(J,U)$ such that the mild solution $x(\cdot)$ of (1) satisfies $x(b) = x_1$ and $x'(b) = y_1$.

To establish our results, we introduce the following assumptions on system (1):

(H_1) $f : J \times X \times X \to X$ is a continuous function and there exist positive constants k_1 and k_2 such that $\|f(t,x_1,y_1) - f(t,x_2,y_2)\| \le k_1\|x_1 - x_2\| + k_2\|y_1 - y_2\|$ for every $x_1,\ x_2,\ y_1$ and $y_2 \in X$.

(H_2) The functions $I_k^j : X \to X$ are continuous and there exist positive constants $L(I_k^j),\ j = 1,2$ such that $\|I_k^j(x_1) - I_k^j(x_2)\| \le L(I_k^j)\|x_1 - x_2\|$ for each $x_1, x_2 \in X$.

(H_3) The linear operator $G_1 : L^2(J,U) \to X$ defined by $G_1u = \int_0^b S(b-s)Bu(s)\mathrm{ds}$ and there exists abounded inverse operator $G_1^{-1} : L^2(J,U)/\mathrm{ker}G_1 \to X$ and a positive constant M_1 such that $\|G_1^{-1}\| \le M_1$.

(H_4) The linear operator $G_2 : L^2(J,U) \to X$ defined by $G_2u = \int_0^b C(b-s)Bu(s)\mathrm{ds}$ and there exists abounded inverse operator $G_2^{-1} : L^2(J,U)/\mathrm{ker}G_2 \to X$ and a positive constant M_2 such that $\|G_2^{-1}\| \le M_2$.

(H_5) $G_1G_2^{-1} = G_2G_1^{-1} = 0$.

(H_6) Let $\mu_1 = (N + M)\widetilde{\beta}k_1b + (M + N\widetilde{\eta} + \widetilde{N} + M\widetilde{\eta})\sum_{k=1}^m L(I_k^1)$ and $\mu_2 = (N + M)\widetilde{\beta}(k_2b + \sum_{k=1}^m L(I_k^2))$, where $\widetilde{\beta} = 1 + Kb(M_1N + M_2M)$, $\widetilde{\eta} = Kb(M_1M + M_2\widetilde{N})$, $\widetilde{N} = \sup_{t\in J}\|AS(t)\|_{L(E,X)}$.

3 Controllability Result

Theorem 1 *If the conditions* $(H_1) - (H_6)$ *and* $\max\{\mu_1, \mu_2\} < 1$ *are satisfied, then the second order impulsive system (1) is exactly controllable on* J.

Proof Let $\|B\| \le K$. In order to prove the exact controllability result, we define the control function by

$$\begin{aligned} u(t) = {} & G_1^{-1}[x_1 - C(b)x_0 - S(b)y_0 - \int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{d}s \\ & - \sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+))](t) \\ & + G_2^{-1}[y_1 - AS(b)x_0 - C(b)y_0 - \int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s \\ & - \sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+))](t) \end{aligned} \tag{3}$$

Let $Z = PC \times PC$ be the space $Z = \{(x,z) : x,z \in PC(J,X)$ and $x'(t) = z(t)$ for $t \neq t_k\}$ provided with the norm $\|(x,z)\|_Z = \max\{\|x\|_{PC}, \|z\|_{PC}\}$. For $(x,z) \in Z$, define the nonlinear operator $\Phi(x,z) = (\Phi_1(x,z), \Phi_2(x,z))$, where

$$\begin{aligned}
&\Phi_1(x,z)(t)\\
&= C(t)x_0 + S(t)y_0 + \textstyle\int_0^t S(t-s)f(s,x(s),x'(s))\mathrm{ds} + \int_0^t S(t-s)B\\
&\quad \cdot\{G_1^{-1}[x_1 - C(b)x_0 - S(b)y_0 - \textstyle\int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+))](s)\\
&\quad +G_2^{-1}[y_1 - AS(b)x_0 - C(b)y_0 - \textstyle\int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+))](s)\}\mathrm{ds}\\
&+\sum_{0<t_k<t} C(t-t_k)I_k^1(x(t_k)) + \sum_{0<t_k<t} S(t-t_k)I_k^2(x'(t_k^+)),
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\Phi_2(x,z)(t)\\
&= AS(t)x_0 + C(t)y_0 + \textstyle\int_0^t C(t-s)f(s,x(s),x'(s))\mathrm{ds} + \int_0^t C(t-s)B\\
&\quad \cdot\{G_1^{-1}[x_1 - C(b)x_0 - S(b)y_0 - \textstyle\int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+))](s)\\
&\quad +G_2^{-1}[y_1 - AS(b)x_0 - C(b)y_0 - \textstyle\int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+))](s)\}\mathrm{ds}\\
&+\sum_{0<t_k<t} AS(t-t_k)I_k^1(x(t_k)) + \sum_{0<t_k<t} C(t-t_k)I_k^2(x'(t_k^+)).
\end{aligned} \tag{5}$$

Substituting the control (3) in (4) and using the hypothesis (H_5), we get

$$\begin{aligned}
&\Phi_1(x,z)(b)\\
&= C(b)x_0 + S(b)y_0 + \textstyle\int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{ds} + \int_0^b S(b-s)B\\
&\quad \cdot\{G_1^{-1}[x_1 - C(b)x_0 - S(b)y_0 - \textstyle\int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+))](s)\\
&\quad +G_2^{-1}[y_1 - AS(b)x_0 - C(b)y_0 - \textstyle\int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad -\sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+))](s)\}\mathrm{ds}\\
&+\sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) + \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+)) = x_1
\end{aligned}$$

and from (5) we get

$$\begin{aligned}
&\Phi_2(x,z)(b)\\
&= AS(b)x_0 + C(b)y_0 + \int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s + \int_0^b C(b-s)B\\
&\quad \cdot\{G_1^{-1}[x_1 - C(b)x_0 - S(b)y_0 - \int_0^b S(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad - \sum_{k=1}^m C(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m S(b-t_k)I_k^2(x'(t_k^+))](s)\\
&\quad + G_2^{-1}[y_1 - AS(b)x_0 - C(b)y_0 - \int_0^b C(b-s)f(s,x(s),x'(s))\mathrm{d}s\\
&\qquad - \sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) - \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+))](s)\}\mathrm{d}s\\
&+ \sum_{k=1}^m AS(b-t_k)I_k^1(x(t_k)) + \sum_{k=1}^m C(b-t_k)I_k^2(x'(t_k^+)) = y_1.
\end{aligned}$$

Hence, the system (1) is controllable provided the operator Φ has a fixed point in Z. The proof is based on the classical fixed point theorem for contractions. It follows from the assumptions that each $\Phi_i,\ i = 1,2$ is well defined and continuous. In order to prove that Φ is a contraction mapping on Z, we take $(x,z),(v,w) \in Z$. From the conditions $(H_1) - (H_4)$, we get

$$\begin{aligned}
&\|\Phi_1(x,z)(t) - \Phi_1(v,w)(t)\|\\
&\le \int_0^b N(k_1\|x-v\| + k_2\|x'-v'\|)\mathrm{d}s\\
&\quad + NKb\{M_1[\int_0^b N(k_1\|x-v\| + k_2\|x'-v'\|)\mathrm{d}s\\
&\qquad + \sum_{k=1}^m ML(I_k^1)\|x(t_k) - v(t_k)\| + \sum_{k=1}^m NL(I_k^2)\|x'(t_k^+) - v'(t_k^+)\|]\\
&\qquad + M_2[\int_0^b M(k_1\|x_s - v_s\| + k_2\|x'_s - v'_s\|)\mathrm{d}s\\
&\qquad + \sum_{k=1}^m \widetilde{N}L(I_k^1)\|x(t_k) - v(t_k)\| + \sum_{k=1}^m ML(I_k^2)\|x'(t_k^+) - v'(t_k^+)\|]\} \qquad (6)\\
&\quad + \sum_{k=1}^m ML(I_k^1)\|x(t_k) - v(t_k)\| + \sum_{k=1}^m NL(I_k^2)\|x'(t_k^+) - v'(t_k^+)\|\\
&\le \{N[1 + Kb(M_1N + M_2M)]k_1b +\\
&\quad [M + NKb(M_1M + M_2\widetilde{N})]\sum_{k=1}^m L(I_k^1)\}\|x - v\|\\
&\quad + N[1 + Kb(M_1N + M_2M)][k_2b + \sum_{k=1}^m L(I_k^2)]\|z - w\|.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\|\Phi_2(x,z)(t)-\Phi_2(v,w)(t)\| \\
&\le \int_0^b M(k_1\|x-v\|+k_2\|x'-v'\|)\mathrm{ds} \\
&\quad +MKb\{M_1[\int_0^b N(k_1\|x-v\|+k_2\|x'-v'\|)\mathrm{ds} \\
&\qquad +\sum_{k=1}^m ML(I_k^1)\|x(t_k)-v(t_k)\|+\sum_{k=1}^m NL(I_k^2)\|x'(t_k^+)-v'(t_k^+)\|] \\
&\qquad +M_2[\int_0^b M(k_1\|x-v\|+k_2\|x'-v'\|)\mathrm{ds} \\
&\qquad +\sum_{k=1}^m \widetilde{N}L(I_k^1)\|x(t_k)-v(t_k)\|+\sum_{k=1}^m ML(I_k^2)\|x'(t_k^+)-v'(t_k^+)\|]\} \\
&\quad +\sum_{k=1}^m \widetilde{N}L(I_k^1)\|\widetilde{x}(t_k)-\widetilde{v}(t_k)\|+\sum_{k=1}^m ML(I_k^2)\|\widetilde{x}'(t_k^+)-\widetilde{v}'(t_k^+)\| \\
&\le \{M[1+Kb(M_1N+M_2M)]k_1b+ \\
&\quad [\widetilde{N}+MKb(M_1M+M_2\widetilde{N})]\sum_{k=1}^m L(I_k^1)\}\|x-v\| \\
&\quad +M[1+Kb(M_1N+M_2M)][k_2b+\sum_{k=1}^m L(I_k^2)]\|z-w\|.
\end{aligned} \tag{7}$$

The above inequalities (6) and (7) and the assumption $\max\{\mu_1,\mu_2\}<1$ imply that Φ is a contraction mapping. Hence there exists a unique fixed point $(x,z)\in Z$. Then function x in $PC^1(J,X)$ is a mild solution of (1). Thus, system (1) is exactly controllable. □

Acknowledgement The authors wish to thank the anonymous referees for their comments and suggestions. This work is supported by NSF of China (No. 11371087) and NSF of Shanghai (No. 12ZR1400100).

References

1. Bainov, D.D., Simeonov, P.S.: System with Impulse Effect. Horwood, Chichester (1989)
2. Balachandran, K., Kim, J.H.: Remarks on the paper "Controllability of the second order differential inclusion in Banach spaces" [J. Math. Anal. Appl. **285,** 537–550 (2003)]. J. Math. Anal. Appl. **324,** 746–749 (2006)
3. Chang, Y.K., Li, W.T.: Controllability of second order differential and integrodifferential inclusions in Banach spaces. J. Optim. Theo. Appl. **129,** 77–87 (2006)
4. Gorniewicz, L., Ntouyas, S.K., Regan, D.O.: Existence and controllability results for first- and second-order functional semilinear differential inclusions with nonlocal conditions. Numer. Funct. Anal. Optim. **28,** 53–82 (2007)
5. Kang, J.R., Kwun, Y.C., Park, J.Y.: Controllability of the second order differential inclusion in Banach spaces. J. Math. Anal. Appl. **285,** 537–550 (2003)
6. Michel, A.N., Hou, L., Liu, D.: Stability of Dynamical Systems, Continuous, Discontinuous and Discrete Systems. Birkhauser, Boston (2008)
7. Park, J.Y., Park, S.H., Kang, Y.H.: Controllability of second-order impulsive neutral functional differential inclusions in Banach spaces. Math. Methods Appl. Sci. **33,** 249–262 (2010)

8. Sakthivel, R., Mahmudov, N.I., Kim, J.H.: On controllability of second order nonlinear impulsive differential systems. Nonlinear. Anal. **71,** 45–52 (2009)
9. Travis, C.C., Webb, G.F.: Compactness, regularity, and uniform continuity properties of strongly continuous cosine families. Houst. J. Math. **3,** 555–567 (1977)
10. Travis, C.C., Webb, G.F.: Cosine families and abstract nonlinear second order differential equations. Acta Math. Acad. Sci. Hung. **32,** 76–96 (1978)

SIAC Filtering for Nonlinear Hyperbolic Equations

Xiaozhou Li and Jennifer K. Ryan

Abstract We present the results of the symmetric and one-sided smoothness-increasing accuracy-conserving (SIAC) filter applied to a discontinuous Galerkin (DG) approximation for two examples of nonlinear hyperbolic conservation laws. The traditional symmetric SIAC filter relies on having a translation invariant mesh, periodic boundary conditions, and linear equations. However, for practical applications that are modeled by nonlinear hyperbolic equations, this is not feasible. Instead we must concentrate on a filter that allows error reduction for nonuniform/unstructured meshes and nonperiodic boundary conditions for nonlinear hyperbolic equations. This proceedings is an introductory exploration into the feasibility of these requirements for efficient filtering of nonlinear equations.

1 Introduction and Motivation

In this chapter, we consider the usefulness of superconvergence extraction techniques for discontinuous Galerkin (DG) approximations to nonlinear hyperbolic equations of the form

$$u_t + \sum_{i=1}^{d} f(u)_{x_i} = 0, (\mathbf{x}, t) \in \Omega \times (0, T], \tag{1}$$

$$u(\mathbf{x}, 0) = u_o(\mathbf{x}), \mathbf{x} \in \Omega. \tag{2}$$

The specific extraction technique that we consider is smoothness-increasing accuracy-conserving (SIAC) filtering. We consider this technique as it is known for

X. Li (✉)
Delft Institute of Applied Mathematics, Delft University of Technology,
2628CD Delft, Netherlands
e-mail: X.Li-2@tudelft.nl

J. K. Ryan
School of Mathematics, University of East Anglia, Norwich NR4 7TJ, UK
e-mail: Jennifer.Ryan@uea.ac.uk

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_41

reducing the oscillations in the DG error as well as the error itself, while increasing the continuity of the numerical approximation.

Mathematically, the symmetric SIAC filter relies on having a translation invariant mesh, periodic boundary conditions, and a linear equation. However, for practical applications that are modeled by nonlinear hyperbolic equations, this is not feasible. Instead we must concentrate on a filter that allows error reduction for nonuniform/unstructured meshes and nonperiodic boundary conditions for nonlinear hyperbolic equations. The question we seek to answer is how feasible are these requirements for efficient filtering of nonlinear equations.

2 Background

2.1 Discontinuous Galerkin Methods

In this section, we merely summarize the important properties of the discontinuous Galerkin method that are useful in superconvergence extraction. More on these methods can be found in [2, 3].

The useful properties are:

- An approximation space that consists of piecewise polynomials of degree $\leq k$
- Weak continuity at element interfaces
- A variational formulation

$$((u_h)_t, \psi)_\Omega + \sum_{i=1}^{d} \left(-(f_i(u_h), \psi_{x_i})_\Omega + \sum_K \int_{\partial K} \widehat{f}_i(u_h^L, u_h^R)\nu_i \psi \, ds \right) = 0, \quad (3)$$

where the summation is over all elements in our discretized domain.

These properties allow us to obtain the following error estimates for the DG solution for linear hyperbolic equations:

- $u - u_h \sim \mathcal{O}(h^{k+1})$ in L^2 for sufficiently smooth initial data, u_0
- $u - u_h \sim \mathcal{O}(h^{2k+1})$ in a negative order norm

We emphasize that these estimates rely on having smooth enough initial data and a linear equation so that information propagates along characteristics. In the case of nonlinear hyperbolic equations, the initial data may be smooth enough, but characteristics may intersect, forming a shock.

2.2 Smoothness-Increasing Accuracy-Conserving (SIAC) Filtering

The SIAC filter is a form of superconvergence extraction that filters out oscillations in the error. It is performed by convolving the DG solution with a B-spline kernel at the final time,

$$u_h^\star(x) = (K_h^{2(k+1),k+1} \star u_h(\cdot, T))(x). \tag{4}$$

Using a SIAC filter on linear hyperbolic equations, one may show that

$$\|u - K_h^{2(k+1),k+1} \star u_h\|_{0,\Omega_0} \leq C\, h^{2k+1}, \tag{5}$$

for a translation invariant mesh. This is based upon the works of [1, 4, 7, 8].

The symmetric convolution kernel is a central B-spline kernel given by

$$K^{(r+1,\ell)}(x) = \sum_{\gamma=0}^{r} c_\gamma^{(r+1,\ell)} \psi^{(\ell)}\left(x - x_\gamma\right), \tag{6}$$

where $K_H^{(r+1,\ell)}(x) = \frac{1}{H} K^{(r+1,\ell)}\left(\frac{x}{H}\right)$, $x_\gamma = -\frac{r}{2} + \gamma$, and generally $r = 2k$ and $\ell = k + 1$. We note that $\psi^{(\ell)}$ is a central B-spline of order ℓ, and H is generally the translation invariance of the mesh. The weighting coefficients of the B-splines are given by the linear system

$$K^{(r+1,\ell)} \star p = p, \qquad p = 1, x^2, \ldots, x^r. \tag{7}$$

Note that convolving the DG solution with such a kernel produces an approximation that is a polynomial of degree $r + 1 \leq 2k + 1$ with continuity of $\ell - 2 \leq k - 1$. Further, note that the postprocessing stencil width is of length $(r + \ell)H$.

2.3 Boundary Filtering

The kernel given in Eq. (6) is for postprocessing smooth regions, away from boundaries. However, when near a boundary or discontinuity, this needs to be sufficiently modified to balance accuracy constraints with error reduction and computational efficiency. It has recently been shown [6] that a suitable modification is given by

$$K^{(r+1,\ell)}(x) = \underbrace{\sum_{\gamma=0}^{r} c_\gamma^{(\ell)} \psi^{(\ell)}(x - x_\gamma - \lambda(x))}_{\text{Shifted filter}} + \underbrace{c_{r+1}^{(\ell)}(x - (\bar{x} - 1))^{\ell-1} \chi_{[\bar{x}-1,\bar{x}]}}_{\text{Special B-spline}}. \tag{8}$$

This kernel uses $r + 1$ central B-splines that are shifted to accommodate a nonsymmetric support near a boundary or discontinuity along with a general B-spline that aids in improving the computational efficiency and reducing the errors in regions where one-sided filters are necessary.

At the price of computational efficiency and error reduction, we have had to give up the property of superconvergence for $\ell - 1 \geq 2$. In the interior, we achieve superconvergence of order $r + 1 \leq 2k + 1$, but still only have convergence of order

$\ell - 1 \leq k + 1$ at the boundaries. However, in the case of linear approximations and $k = 1$, we still achieve a global superconvergence order of three, *even in the boundary regions.* This clearly shifts our focus to error reduction and hence allows us to more closely examine how the SIAC filter could aid in error reduction for nonlinear equations whose solution contains a discontinuity.

3 SIAC Filtering for Nonlinear Hyperbolic Equations

There has been previous work in SIAC filtering for nonlinear hyperbolic equations. However, the work was restricted to nonlinear equations with a *smooth* solution. In [5], the following theorem was given:

Theorem 1 *Assume we have a smooth solution to Eq. (1) whose DG approximate is given by u_h. If $|f_i''| \leq M$, then*

$$\|(u - u_h)(T)\|_{-(k+1),\Omega} \leq Ch^{2k+m}, \tag{9}$$

where $m = 0, \frac{1}{2}, 1$, depends on the numerical flux and $k > \frac{d}{2}$.
As a consequence of this higher order convergence in the negative-order norm, we then have $\mathcal{O}(h^{2k+m})$ convergence of the postprocessed solution in the L^2-norm.

4 Numerical Examples

Although the theory has been established for smooth solutions, it is interesting to investigate the application of the SIAC filter to nonsmooth solutions. To demonstrate the possibilities of the SIAC filtered solution for such solutions, we present two examples: First, a one-dimensional Burgers equation after the shock has developed; and second, the double Mach reflection problem of the two-dimensional Euler equations.

The steps of the filtering process are as follows:

- Calculate the DG approximation to the equation at the final time $t = T$
- Identify "troubled cells," i.e., where there is a discontinuity
- Calculate the SIAC filtered solution
  - Use a symmetric filter in smooth regions, a distance of at least $\frac{r+\ell}{2}h$ away from boundaries or discontinuities
  - In "troubled cell regions," use a boundary filter

4.1 One-Dimensional Burgers Equation

For the first example, we consider the equation

$$u_t + uu_x = 0, \quad u(x,0) = \sin(x), \qquad x \in [0, 2\pi], \quad T = 1. \tag{10}$$

Table 1 The L^2-error of the DG solution and the SIAC filtered DG solution for $\mathbb{P}^2$ and $\mathbb{P}^3$ using the boundary filter in the appropriate regions. Errors are calculated away from the shock

	DG		SIAC DG		DG		SIAC DG	
Mesh	L^2 error	order	L^2 error	order	L^2 error	order	L^2 error	order
	$\mathbb{P}^2$				$\mathbb{P}^3$			
40	1.02E-05	–	5.28E-06	–	4.89E-08	–	5.18E-08	–
60	3.29E-06	2.80	1.61E-06	2.94	1.07E-08	3.74	1.59E-09	8.60
80	1.46E-06	2.83	6.94E-07	2.92	3.08E-09	4.33	1.61E-10	7.96
100	7.84E-07	2.78	3.63E-07	2.90	1.33E-09	3.76	3.40E-11	6.96

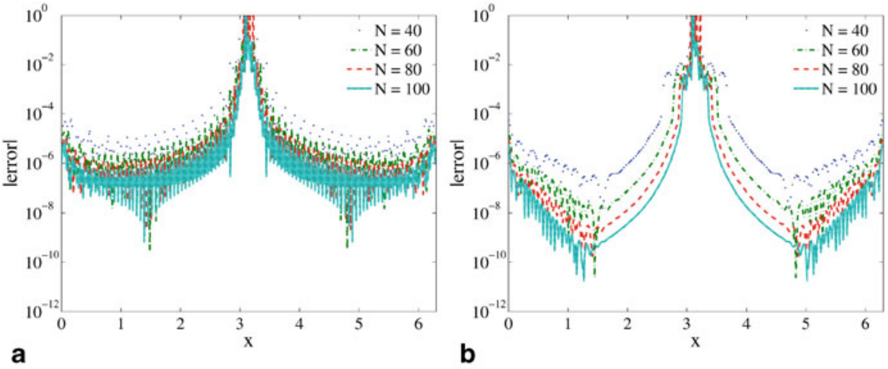


Fig. 1 Plots of pointwise errors of the DG solution and the SIAC filtered DG solution for $\mathbb{P}^2$ using the boundary filter in the appropriate regions

Note that this equation contains a shock at $x = \pi$. We have implemented the symmetric filter in smooth regions and the boundary filter in the elements next to the boundaries and shocks. No filter is implemented in the element that contains the shock. The results for the errors are presented in Table 1 and Fig. 1.

4.2 Two-Dimensional Double Mach Reflection

In this example, we apply the SIAC filter, including the boundary filter, to the two-dimensional Euler equations for the double Mach reflection problem. We use the multiwavelet-troubled cell indicator of Vuik [9] and plot the results for a zoomed-in region of the solution in Fig. 2. Note that from the results given for Burgers equation, we expect that the difference when we examine the two solutions will be small. However, we do observe some reduced oscillations with the SIAC filtered DG approximation.

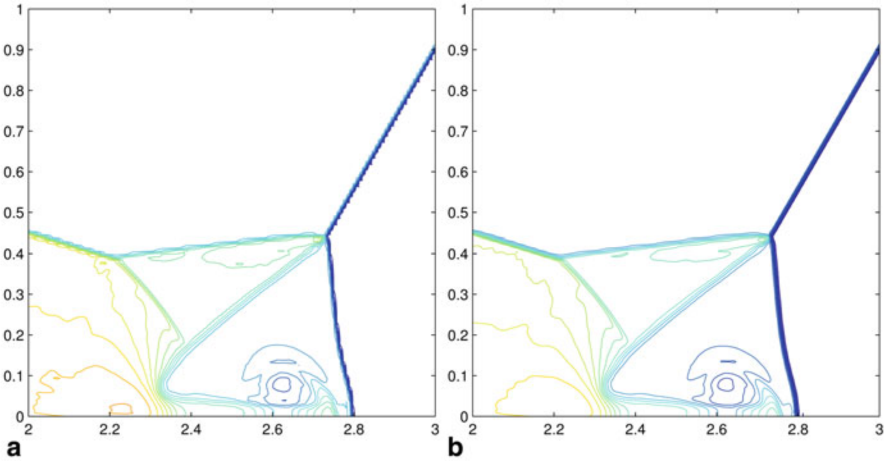


Fig. 2 Results for the DG approximation and SIAC filtered DG approximation when applied to the double Mach reflection problem

5 Conclusions and Future Work

SIAC filtering holds promise in applications to nonlinear equations, although their exact usefulness remains unclear. Traditionally, SIAC filtering uses B-splines to induce smoothness on the DG field and enhance accuracy. This traditionally allows order improvement from $\mathcal{O}(h^{k+1})$ to $\mathcal{O}(h^{2k+m})$ for smooth regions. At the boundaries, order is reduced for improved computational efficiency. For nonlinear equations, their usefulness depends on the boundedness of the flux function and the chosen numerical flux. From our observations, the filtering appears to reduce oscillations in regions where applied. How exactly it should be applied is the subject of on-going research.

Acknowledgement Portions of this research are supported by the European Office of Aerospace Research and Development under grant numbers FA8655-09-1-3055 and FA8655-13-1-3017. We would like to thank Liangyue Ji (Minnesota) for useful discussions and Thea Vuik (TU Delft) for providing the data used in the double Mach reflection example.

References

1. Bramble, J.H., Schatz, A.H.: Higher order local accuracy by averaging in the finite element method. Math. Comput. **31,** 94–111 (1977)
2. Cockburn, B.: Discontinuous Galerkin methods for convection-dominated problems. High-Order Methods for Computational Physics, vol. 9 of Lecture Notes in Computational Science and Engineering. Springer (1999)
3. Cockburn, B., Shu, C.-W.: Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. J. Sci. Comput. **16**, 173–261 (2001)

4. Cockburn, B., Luskin, M., Shu, C.-W., Süli, E.: Enhanced accuracy by post-processing for finite element methods for hyperbolic equations. Math. Comput. **72**, 577–606 (2003)
5. Ji, L., Xu, Y., Ryan, J.K.: Negative-order norm estimates for nonlinear hyperbolic conservation laws. J. Sci. Comput. **54**, 269–310 (2013)
6. Ryan, J.K., Li, X., Kirby, R.M., Vuik, C.: One-sided position-dependent smoothness-increasing accuracy-conserving (SIAC) filtering over uniform and non-uniform meshes. J. Sci. Comput., accepted.
7. Mock, M.S., Lax, P.D.: The computation of discontinuous solutions of linear hyperbolic equations. Commun. Pure Appl. Math. **18**, 423–430 (1978)
8. Thomée, V.: High order local approximations to derivatives in the finite element method. Math. Comput. **31**, 652–660 (1977)
9. Vuik, M.J.: Limiting and shock detection for discontinuous Galerkin solutions using multiwavelets. TU Delft MSc Thesis, 2012-08-24

Structural Analysis and Dummy Derivatives: Some Relations

R. McKenzie and J. Pryce

Abstract Differential algebraic equations (DAEs) appear frequently in applications involving equation based modeling, from robotics to chemical engineering. A common way of making a DAE amenable to numerical solution is by reducing the index to obtain a corresponding ordinary differential equations (ODE) and using an ODE solution method. The signature matrix method developed by Pryce does not rely on an index reduction step and instead solves the DAE directly via Taylor series. The chapter draws comparisons between these two different approaches and shows the signature matrix method is in some sense equivalent to the dummy derivative index reduction method developed by Mattsson and Söderlind. The ideas are illustrated via a DAE from Campbell and Griepentrog that models a robot arm. The authors acknowledge G. Tan and N. Nedialkov at McMaster University, Hamilton, Canada for their support in this chapter and the talk that accompanied it at AMMCS-2013.

1 An Overview of the Structural Analysis Method

We present a short explanation of the structural analysis (SA) method, which is used in Sect. 3 to compare the SA approach and dummy derivative (DD) approach.

We are given a system of n equations $f_i = 0$ in n unknown functions x_j of time t, where the equations may contain derivatives of the n unknowns. We form the problem's signature matrix $\varSigma$, with entries of the form:

$$\sigma_{ij} = \begin{cases} \text{order of highest derivative of } x_j \text{ in } f_i & \text{if } x_j \text{ occurs in } f_i, \\ -\infty & \text{if not}\,. \end{cases} \tag{1}$$

R. McKenzie (✉) · J. Pryce
Cardiff University, Cardiff, UK
e-mail: mckenzier1@cardiff.ac.uk

J. Pryce
e-mail: j.d.pryce@cantab.net

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_42

We then find a highest value transversal (HVT) for our problem. This is found by taking one finite entry in each row and column of Σ such that the total sum is greater than that of any other choice of finite entry in each row and column. If this is possible the problem is called structurally well-posed (SWP); if not it is called structurally ill-posed (SIP).

We look for nonnegative integer valued offset vectors, **c** and **d** satisfying:

$$\sigma_{ij} \leq d_j - c_i, \tag{2}$$

with equality on a HVT and $\min_i c_i = 0$.
We then form the $n \times n$ Jacobian matrix **J** with entries

$$\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j^{(d_j - c_i)}} = \begin{cases} \frac{\partial f_i}{\partial x_j^{(\sigma_{ij})}} & \text{if } d_j - c_i = \sigma_{ij} \\ 0 & \text{elsewhere} \end{cases}$$

and solve the equations to obtain Taylor coefficients in steps numbered as k, using equations

$$f_i^{(k+c_i)} = 0 \qquad \forall i \text{ such that } k + c_i \geq 0 \tag{3}$$

to solve for variables

$$x_j^{(k+d_j)} \qquad \forall j \text{ such that } k + d_j \geq 0 \tag{4}$$

at each step. The initial step number is equal to $-\max_j d_j$. If rows and columns of **J** are ordered by descending offset values, then at each stage of the solution process the Jacobian ($\mathbf{J}_k$) is a submatrix of **J**, the Jacobian at stage 0, see [6].

2 An Overview of the DD Method

We define some notation used for the DD method in [3]. Write the original problem as $\mathcal{F}x = 0$, where $\mathcal{F}$ is a (column n-vector) differential–algebraic operator (DAO).

1. $\nu(\mathcal{F})$, a column n-vector of nonnegative integers, containing the minimum number of differentiations of each equation to derive an ODE.
2. $D^{\nu} = \text{diag}\left(\frac{d^{\nu}}{dt^{\nu}}, \ldots, \frac{d^{\nu_n}}{dt^{\nu_n}}\right)$, regarded as a DAO.
3. The differentiated problem is $\mathcal{G}x = D^{\nu(\mathcal{F})}\mathcal{F}x = 0$.

We can now present the main steps of the dummy derivative algorithm:

1. Find $\nu(\mathcal{F})$ (by Pantelides' method or SA).
2. Obtain a differentiated problem $\mathcal{G}x = 0$.
3. Permute the Jacobian of $\mathcal{G}x$ to block lower-triangular form.
4. Perform the index reduction algorithm. Loop through steps that select derivatives to be considered as algebraic variables in the solution process.

We need to make several assumptions and define a way of indexing over blocks before coming to the main index reduction step (item 4) of the algorithm. Let g_i represent the ith diagonal block in $\mathcal{G}$, and let z_i be the vector of highest-order derivatives (HODs) of block g_i. Assume that $\mathcal{G}$ is in block lower-triangular form with only one block. Also assume the equations (and variables) have been sorted into descending order with respect to number of differentiations. We consider each step in turn by the superscript $[j]$. The index reduction part of the algorithm is:

1. Initialize $z^{[0]}$, $g^{[0]}(z^{[0]})$, $j = 0$, $G^{[0]} = \dfrac{\partial g^{[0]}}{\partial z^{[0]}}$.
2. While $g^{[j]}$ has m differentiated equations, with $m > 0$:
3. Let $H^{[j]}$ be the first m rows of $G^{[j]}$.
4. Choose m columns of $H^{[j]}$ to get a square nonsingular matrix $G^{[j+1]}$.
5. Set the corresponding derivatives of the variables considered in $G^{[j+1]}$ to be dummy derivatives.
6. Omit one differentiation.
7. Set $j = j + 1$.
8. End while.

Now collect all original and differentiated equations used, in original variables and dummy derivatives, to get a new square system of index 1.

3 A Comparison Between the Two Methods

There are many similarities between the two methods. First, $\nu(\mathcal{F})$ found in DDs is the same as $\mathbf{c}$ in SA. This is apparent because from [6] we see that Pantelides' algorithm [4] and SA can be used interchangeably. Therefore we have that $D^{\nu} = \operatorname{diag}\left(\dfrac{d^{c_1}}{dt^{c_1}}, \ldots, \dfrac{d^{c_n}}{dt^{c_n}}\right)$. Now we have the following equalities:

$$\mathcal{G}x = \mathcal{F}^{\nu}x = D^{\nu(\mathcal{F})}\mathcal{F}x = D^{c}\mathcal{F}x, \tag{5}$$

and thus the differentiated problem can be written as $(f_i^{(c_i)}(x) = 0)_{i=1}^{n}$. Hence, the first stage system in DDs is the $k = 0$ stage system in SA. We are differentiating the ith equation c_i times, so the maximum derivative for each variable x_j in $\mathcal{G}x = 0$ will be equal to $\max_i(\sigma_{ij} + c_i)$. From (2) this is d_j. Hence we have $z^{[1]} = (x_1^{(d_1)}, \ldots, x_n^{(d_n)})$.

The formula for the DD Jacobian matrix $G^{[0]}$ can now be written in this SA-based notation to show it equals the SA Jacobian $\mathbf{J}$.

$$G^{[0]} = \frac{\partial g_i^{[1]}}{\partial z_i^{[1]}} = \frac{\partial f_i^{(c_i)}}{\partial x_j^{(d_j)}} = \mathbf{J}. \tag{6}$$

We note that going to the next step in DDs by reducing the order of differentiation by one is equivalent to reducing the offset vector $\mathbf{c}$ by 1 in positions where it is > 0

(and consequently reducing **d** by 1 correspondingly). Therefore at step 2 in DDs we consider the equations used in step -1 of SA, since SA increases the order of differentiation by one at each step. Because of this we renumber the DD method for the remainder of this section to be $0, 1, \ldots$. A proof of both methods using the same equations at each step follows.

Proof We have already shown that at stage 0 both methods use the same equations. In DDs we now remove any equation such that $c_i = 0$. We then omit one differentiation and repeat. Hence at step 1 we remove equations such that $c_i - 1 = 0$ and so on. That is, at subsequent step k remove equations such that $c_i - k = 0$. From (3) this gives exactly the equations considered at step $-k$ in SA. □

Before we go in to any deeper comparisons we need the following lemma.

Lemma 1 *If a square nonsingular DAE has* 0 *D.O.F. and a HVT on the main diagonal of Σ then $d = c^T$.*

Proof If we have 0 D.O.F. this is equivalent to $\sum_j d_j - \sum_i c_i = 0$ from [6] and by (2) $d_i - c_i \geq 0,\ \forall i$. Hence, $d_i \geq c_i$ and so both must be equal. □

Let us now assume we have 0 D.O.F. and have organized the BTF so that there is a HVT on the main diagonal. We already have $G^{[0]} = \mathbf{J}_0$. Let us prove this equality extends to further steps. Let m_k be the number of variables at the end of each step (the number of DDs). At step k, the nonsquare matrix $H^{[k]}$ has m_k equations in m_{k-1} variables. The $m_{k-1} - m_k$ equations removed in going from $G^{[k]}$ to $H^{[k]}$ all had $c_i - k = 0$, so from the lemma there must also be $m_{k-1} - m_k$ variables with $d_j - k = 0$. If $1 \leq i \leq m_k$ and $j > m_k$ then $-k + c_i > 0$ and $-k + d_j \leq 0$, hence $d_j - c_i < 0$ and thus cannot be equal to σ_{ij}, so that $\mathbf{J}_{-k,ij} = 0$. Since we have equality at the initial step in DDs this means $G^{[0]}_{ij} = 0$ and by induction $G^{[k]}_{ij} = 0$. Thus columns with $d_j = 0$ must be removed to form $G^{[k+1]}$ for structural reasons, i.e., we do not consider undifferentiated variables. This leaves us looking at variables such that $-k + d_j > 0$ at each stage, which is equivalent to all differentiated variables solved for in SA at the equivalent step, i.e., $G^{[k]} = \mathbf{J}_{-k}$.

Unfortunately we cannot be as certain as to which variables will become dummy derivatives once degrees of freedom are introduced. Of course, we can always know the total number of dummy derivatives introduced will be $\sum_i c_i$. As to which variables become dummy derivatives, the best we can say at present is the following.

At each step the variables that become dummy derivatives are a subset of the variables that became dummy derivatives at the previous step (necessarily excluding those with $k - d_j = 0$), of size m_k, with each variable being differentiated one time less than in the previous step. So, the dummy derivatives will be a subset of the variables found at the equivalent step of SA. More formally, index the set of dummy derivatives at step k as $D^{[k]}$. Then, regarding variables and their derivatives as distinct symbols

$$D^{[k]} \subset \left\{ x_j^{(-k+d_j)} \mid (-\,k + d_j) > 0, j = 1, \ldots, n \right\} \tag{7}$$

where the $D^{[k]}$ must be chosen in such a way that the set:

$$X^{[k]} = \left\{ x_j \mid x_j^{(l)} \in D^{[k]},\ \text{for some } l \right\} \tag{8}$$

decreases as k increases.

4 An Example

Consider the robot arm DAE, introduced in [1] and reformulated to give a structurally nonsingular Jacobian in [5]. We shall use the 6×6 formulation introduced in [5], where the equations are given in full. For our purposes, it is enough to give its structural Jacobian, signature matrix and offsets (with a HVT marked by $\bullet$ and $-\infty$ entries left blank):

$$\Sigma = \begin{array}{c|cccccc|c}
 & x_1 & x_3 & w & x_2 & u_2 & u_1 & c_i \\
G & 0^{\bullet} & 0 & & & & & 4 \\
H & 0 & 0^{\bullet} & & & & & 4 \\
D & 2 & 1 & 0^{\bullet} & 0 & & & 2 \\
F & 1 & 2 & 0 & 0^{\bullet} & & & 2 \\
E & 1 & 1 & 0 & 2 & 0^{\bullet} & & 0 \\
K & & & 0 & & 0 & 0^{\bullet} & 0 \\
d_j & 4 & 4 & 2 & 2 & 0 & 0 &
\end{array}, \qquad
\mathbf{J} = \begin{array}{c|cccccc|}
 & x_1^{(4)} & x_3^{(4)} & w'' & x_2'' & u_2 & u_1 \\
G^{(4)} & G_{x_1} & G_{x_3} & 0 & 0 & 0 & 0 \\
H^{(4)} & H_{x_1} & H_{x_3} & 0 & 0 & 0 & 0 \\
D'' & D_{x_1''} & 0 & D_w & D_{x_2} & 0 & 0 \\
F'' & 0 & F_{x_3''} & F_w & F_{x_2} & 0 & 0 \\
E & 0 & 0 & 0 & E_{x_2''} & E_{u_2} & 0 \\
K & 0 & 0 & 0 & 0 & K_{u_2} & K_{u_1}
\end{array}.$$

Working through the dummy derivative algorithm yields $\nu(\mathcal{F}) = (4, 4, 2, 2, 0, 0)$, The vector of HODs is $z^{[0]} = (x_1^{(4)}, x_3^{(4)}, w'', x_2'', u_2, u_1)^T$ and $g^{[0]} = (G^{(4)}, H^{(4)}, D^{(2)}, F^{(2)}, E, K)^T$.

Thus we have a dummy derivative Jacobian of the form:

$$\frac{\partial g^{[0]}}{\partial z^{[0]}} = G^{[0]} = \begin{array}{c|cccccc|c}
 & x_1^{(4)} & x_3^{(4)} & w'' & x_2'' & u_2 & u_1 & c_i \\
G^{(4)} & G^{(4)}_{x_1^{(4)}} & G^{(4)}_{x_3^{(4)}} & 0 & 0 & 0 & 0 & 4 \\
H^{(4)} & H^{(4)}_{x_1^{(4)}} & H^{(4)}_{x_3^{(4)}} & 0 & 0 & 0 & 0 & 4 \\
D'' & D''_{x_1^{(4)}} & 0 & D''_{w''} & D''_{x_2''} & 0 & 0 & 2 \\
F'' & 0 & F''_{x_3^{(4)}} & F''_{w''} & F''_{x_2''} & 0 & 0 & 2 \\
E & 0 & 0 & 0 & E_{x_2''} & E_{u_2} & 0 & 0 \\
K & 0 & 0 & 0 & 0 & K_{u_2} & K_{u_1} & 0 \\
d_j & 4 & 4 & 2 & 2 & 0 & 0 &
\end{array}.$$

By Griewank's Lemma [2] this is equivalent to $\mathbf{J}$.

Removing equations with $c_i = 0$ yields the under-determined system:

$$H^{[0]} = \begin{array}{c|cccccc|c} & x_1^{(4)} & x_3^{(4)} & w'' & x_2'' & u_2 & u_1 & c_i \\ G^{(4)} & G^{(4)}_{x_1^{(4)}} & G^{(4)}_{x_3^{(4)}} & 0 & 0 & 0 & 0 & 4 \\ H^{(4)} & H^{(4)}_{x_1^{(4)}} & H^{(4)}_{x_3^{(4)}} & 0 & 0 & 0 & 0 & 4 \\ D'' & D''_{x_1^{(4)}} & 0 & D''_{w''} & D''_{x_2''} & 0 & 0 & 2 \\ F'' & 0 & F''_{x_3^{(4)}} & F''_{w''} & F''_{x_2''} & 0 & 0 & 2 \\ d_j & 4 & 4 & 2 & 2 & 0 & 0 & \end{array} .$$

We are now forced to remove the last two columns to get a nonsingular matrix:

$$G^{[1]} = \begin{array}{c|cccc|c} & x_1^{(4)} & x_3^{(4)} & w'' & x_2'' & c_i \\ G^{(4)} & G^{(4)}_{x_1^{(4)}} & G^{(4)}_{x_3^{(4)}} & 0 & 0 & 4 \\ H^{(4)} & H^{(4)}_{x_1^{(4)}} & H^{(4)}_{x_3^{(4)}} & 0 & 0 & 4 \\ D'' & D''_{x_1^{(4)}} & 0 & D''_{w''} & D''_{x_2''} & 2 \\ F'' & 0 & F''_{x_3^{(4)}} & F''_{w''} & F''_{x_2''} & 2 \\ d_j & 4 & 4 & 2 & 2 & \end{array} .$$

We thus make $x_1^{(4)}$, $x_3^{(4)}$, w'', x_2'' dummy derivatives, reduce the order of differentiation and repeat the process. We end up at the following scheme, with the SA stages listed alongside for comparison:

DD stage	SA stage	Equations being used	Variables being found	DDs selected
4	-4	G, H	x_1, x_3	N/A
3	-3	G', H'	x_1', x_3'	x_1', x_3'
2	-2	G'', H'', D, F	x_1'', x_3'', w, x_2	x_1'', x_3''
1	-1	$G^{(3)}, H^{(3)}, D', F'$	$x_1^{(3)}, x_3^{(3)}, w', x_2'$	$x_1^{(3)}, x_3^{(3)}, w', x_2'$
0	0	$G^{(4)}, H^{(4)}, D'', F'', E, K$	$x_1^{(4)}, x_3^{(4)}, w'', x_2'', u_2, u_1$	$x_1^{(4)}, x_3^{(4)}, w'', x_2''$

The dummy derivatives are equivalent to the differentiated variables solved for at each step in SA as expected, due to the 0 D.O.F. in this example.

5 Conclusions

We have shown the index reduction algorithm in [3] produces a near identical solution scheme to that of SA in some situations, and in any case the SA will provide a set of variables that could be chosen as dummy derivatives at a step. In future work we aim to extend our results to the cases where D.O.F are present.

References

1. Campbell, S.L., Griepentrog, E.: Solvability of general differential algebraic equations. SIAM. J. Sci. Comput. **16**(2), 257–270 (1995)
2. Griewank, A. : Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. (Frontiers in applied mathematics). SIAM, Philadelphia (2000)
3. Mattsson, S.E., Söderlind, G.: Index reduction in differential–algebraic equations using dummy derivatives. SIAM. J. Sci. Comput. **14**(3), 677–692 (1993)
4. Pantelides, C.C.: The consistent initialization of differential–algebraic systems. SIAM. J. Sci. Stat. Comput. **9**, 213–231 (1988)
5. Pryce, J.D.: Solving high-index DAEs by Taylor series. Numer. Algorithms **19**, 195–211 (1998)
6. Pryce, J.D.: A simple structural analysis method for DAEs. BIT. **41**(2), 364–394 (2001)

On the Exact Solutions of the Klein–Gordon–Zakharov Equations

Isaiah Elvis Mhlanga and Chaudry Masood Khalique

Abstract In this chapter we study a coupled system of nonlinear partial differential equations (PDEs), namely, the Klein–Gordon–Zakharov equations. The travelling wave hypothesis approach along with the simplest equation methods are utilized to obtain exact solutions of this system.

1 Introduction

The Klein–Gordon–Zakharov (KGZ) equations [1]

$$u_{texttt} - u_{xx} + u + uv + |u|^2 u = 0, \tag{1a}$$

$$v_{texttt} - v_{xx} - (|u|^2)_{xx} = 0, \tag{1b}$$

are a coupled system of nonlinear partial differential equations (PDEs) by two functions $u(x,t)$ and $v(x,t)$. This model describes the interaction of the Langmuir wave and the ion acoustic wave in plasma. The function $u(x,t)$ denotes the fast time scale component of electric field raised by electrons and the function $v(x,t)$ denotes the deviation of ion density from its equilibrium. Here $u(x,t)$ is a complex function and $v(x,t)$ is a real function. Note that if we remove the term $|u|^2 u$, then this system reduces to the classical Klein–Gordon–Zakharov system regime [2]

$$\begin{aligned} u_{texttt} - u_{xx} + u + uv &= 0, \\ v_{texttt} - v_{xx} - (|u|^2)_{xx} &= 0. \end{aligned} \tag{2}$$

I. E. Mhlanga (✉) · C. M. Khalique
Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, Department of Mathematical Sciences, North-West University, Mafikeng Campus, Private Bag X 2046, Mmabatho 2735, Republic of South Africa
e-mail: Isaiah.Mhlanga@nwu.ac.za

C. M. Khalique
e-mail: Masood.Khalique@nwu.ac.za

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_43

Up to now, a number of studies have been conducted only for this system (2) in different time space [3–7]. Chen Lin [8] considered orbital stability of solitary waves for the KGZ equations (1), while Shi et al. [2] employed the sine–cosine method and the extended tanh method to construct exact wave solutions of the KGZ equations (1).

In this chapter, we employ the travelling wave variable approach along with the simplest equation to obtain exact solutions of the KGZ equations (1).

2 Solution of (1) Using the Travelling Wave Variable Approach

The travelling wave variable approach converts the system of nonlinear PDEs into a nonlinear ordinary differential equation which we then solve to obtain exact solutions of the system.

In order to solve the KGZ equations (1), we first transform it to a system of nonlinear ordinary differential equations which can then be solved in order to obtain its exact solutions.

We make the wave variable transformation

$$u = e^{i\phi}u(z),\ v = v(z),\ \phi = px + rt,\ z = kx + dt, \tag{3}$$

where p, r, k and d are real constants, $d \neq k$. Using this transformation, (1) transforms to

$$(p^2 - r^2 + 1)u + i(2rd - 2pk)u' + (d^2 - k^2)u'' + uv + u^3 = 0, \tag{4a}$$

$$(d^2 - k^2)v'' - (u^2)'' = 0. \tag{4b}$$

Integrating (4b) twice and taking the constants of integration to be zero we obtain

$$v = \frac{u^2}{d^2 - k^2}. \tag{5}$$

Now substituting (5) into (4a) we get

$$u'' = \Big(\frac{r^2 - p^2 - 1}{d^2 - k^2}\Big)u + \Big(\frac{d^2 - k^2 + 1}{(d^2 - k^2)^2}\Big)u^3, \tag{6}$$

which can be written in the form

$$u'' = Au + Bu^3, \tag{7}$$

where
$A = \frac{r^2-p^2-1}{d^2-k^2}$ and $B = \frac{d^2-k^2+1}{(d^2-k^2)^2}$.

Solving (7), with the aid of Mathematica, we obtain the solution

$$u(z) = \pm\frac{1}{P_2}\, i\ \mathrm{sn}(P_1|\omega), \tag{8}$$

where $\mathrm{sn}(P_1|\omega)$ is a Jacobian elliptic function of the *sine-amplitude* [9],

$$P_1 = \frac{\sqrt{\left(\sqrt{A^2-2Bc_1}-A\right)(z+c_2)^{\,2}}}{\sqrt{2}}, \quad P_2 = \sqrt{\frac{B}{\sqrt{A^2-2Bc_1}+A}}$$

and

$$\omega = \frac{-Bc_1 + A\left(\sqrt{A^2-2Bc_1}+A\right)}{Bc_1}$$

is the modulus of the elliptic function with $0 < \omega < 1$. Here c_1 and c_2 are constants of integration. Reverting back to our original variables, we can now write the solution of our Klein–Gordon–Zakharov equations as

$$u(x,t) = \pm\frac{e^{i(px+rt)}}{P_2}\, i\,\mathrm{sn}(P_1|\omega), \tag{9}$$

where

$$P_1 = \frac{\sqrt{\left(\sqrt{A^2-2Bc_1}-A\right)(kx+dt+c_2)^{\,2}}}{\sqrt{2}},$$

ω and P_2 are as above.

Now $v(x,t)$ can be obtained from (5).

It should be noted that the solution (9) is valid for $0 < \omega < 1$ and as ω approaches zero, the solution becomes the normal sine function, $\sin z$, and as ω approaches 1, the solution tends to the tanh function, $\tanh z$.

The profile of the solution (9) is given in the Fig. 1.

3 Solution of (6) Using the Simplest Equation Method

We now use the simplest equation method [10] to solve (6). The simplest equations that will be used are the Bernoulli and Riccati equations.
Let us consider the solutions of (6) in the form

$$u(z) = \sum_{i=0}^{M} \mathcal{A}_i (G(z))^i, \tag{10}$$

where $G(z)$ satisfies the Bernoulli or Riccati equation. M is a positive integer that can be determined by the balancing procedure and $\mathcal{A}_i$, $(i = 0, 1, \cdots, M)$, are parameters to be determined.

We first consider the Bernoulli equation

$$G'(z) = aG(z) + bG^2(z), \tag{11}$$

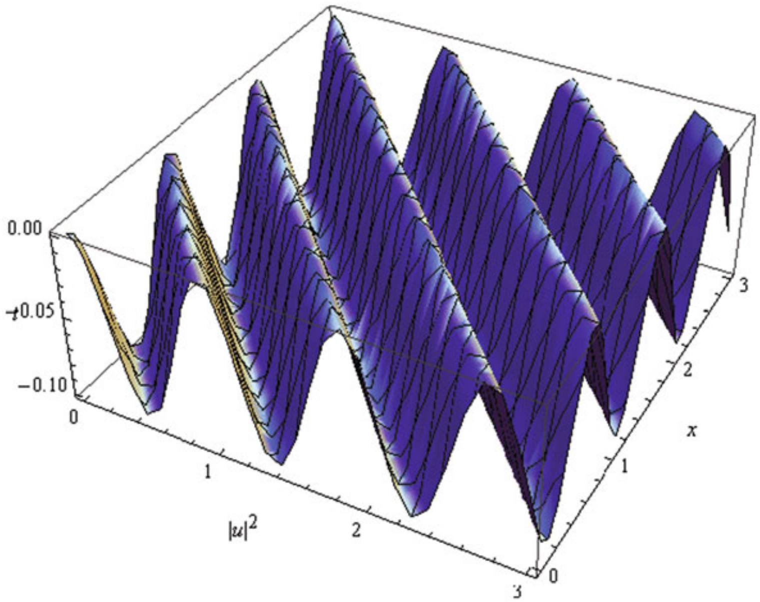


Fig. 1 3D plot of solution (9)

where a and b are constants.

The balancing procedure yields $M = 1$, so the solution of (6) is of the form

$$u(z) = \mathcal{A}_0 + \mathcal{A}_1 G(z). \tag{12}$$

Substituting (12) into (6) and making use of the Bernoulli equation (11) and then equating all coefficients of the function G' to zero, we obtain

$$2\,A_1k^4b^2 - k^2{A_1}^3 + d^2{A_1}^3 - 4\,A_1d^2k^2b^2 + {A_1}^3 + 2\,A_1d^4b^2 = 0,$$

$$3\,A_1k^4ab + 3\,d^2A_0{A_1}^2 - 3\,k^2A_0{A_1}^2 - 6\,A_1d^2k^2ab + 3\,A_0{A_1}^2 + 3\,A_1d^4ab = 0,$$

$$p^2A_0d^2 - p^2A_0k^2 - r^2A_0d^2 + r^2A_0k^2 - A_0k^2 + d^2{A_0}^3 - k^2{A_0}^3 + A_0d^2 + {A_0}^3 = 0,$$

$$A_1d^2 + 3\,d^2{A_0}^2A_1 - p^2A_1k^2 + r^2A_1k^2 + p^2A_1d^2 + A_1d^4a^2 - 3\,k^2{A_0}^2A_1 + A_1k^4a^2$$
$$+\,3\,{A_0}^2A_1 - 2\,A_1d^2k^2a^2 - r^2A_1d^2 - A_1k^2 = 0.$$

Solving this system of algebraic equations with the aid of Maple, we obtain

$$A_0 = \frac{a(d^2-k^2)}{\sqrt{2(k^2-d^2-1)}}, \quad A_1 = \frac{\sqrt{2}b(d^2-k^2)}{k^2-d^2-1},$$

$$p = \sqrt{\frac{r^2d^2 - r^2k^2 + k^2 - d^2{A_0}^2 + k^2{A_0}^2 - d^2 - {A_0}^2}{d^2-k^2}}.$$

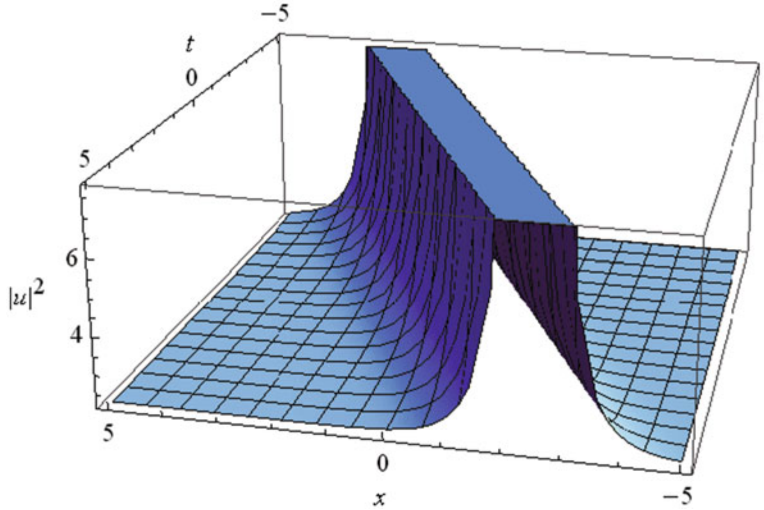


Fig. 2 3D plot of solution (13)

As a result a solution of (1) using the Bernoulli equation as the simplest equation is

$$u(x,t) = e^{i(px+rt)}\Big[\frac{\sqrt{2}ab\left(d^2-k^2\right)(\cosh(a(kx+dt+c))+\sinh(a(kx+dt+c)))}{\sqrt{k^2-d^2-1}(1-b\cosh(a(kx+dt+c))-b\sinh(a(kx+dt+c)))} + \frac{a\left(d^2-k^2\right)}{\sqrt{2(k^2-d^2-1)}}\Big], \tag{13}$$

where c is a constant of integration.

The profile of the solution (13) is given in the Fig. 2.

Similarly for the Riccati equation

$$G'(z) = aG^2(z) + bG(z) + c, \tag{14}$$

where a, b and c are constants.

The balancing procedure yields $M = 1$, so the solution of (6) is of the form

$$u(z) = \mathcal{A}_0 + \mathcal{A}_1 G(z). \tag{15}$$

Similar calculations yield the following set of values

$$A_0 = \frac{b(d^2-k^2)}{\sqrt{2(k^2-d^2-1)}}, \quad A_1 = -\frac{\sqrt{2}a(d^2-k^2)\sqrt{k^2-d^2-1}}{d^2-k^2+1},$$

$$c = -\frac{\sqrt{2(k^2-d^2-1)}(d^2b^2-k^2b^2-2-2p^2+2r^2)}{4A_1(d^2-k^2+1)}.$$

So using Ricatti as the simplest equation, the solutions of the Klein–Gordon–Zakharov equations (1) are

$$u(t,x) = e^{i\phi}\left[\mathcal{A}_0 + \mathcal{A}_1\left\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\left[\frac{1}{2}\theta(z+C)\right]\right\}\right] \tag{16}$$

and

$$\begin{aligned} u(t,\,x) =&e^{i\phi}\Big[\mathcal{A}_0 + \mathcal{A}_1\Big\{-\frac{b}{2a} - \frac{\theta}{2a}\tanh\left(\frac{1}{2}\theta z\right) \\ &+ \frac{\operatorname{sech}\left(\frac{\theta z}{2}\right)}{C\cosh\left(\frac{\theta z}{2}\right) - \frac{2a}{\theta}\sinh\left(\frac{\theta z}{2}\right)}\Big\}\Big], \end{aligned} \tag{17}$$

where $\phi = px + rt$ and $z = kx + dt$. θ is given by $\sqrt{b^2 - 4ac}$, C is a constant of integration, and A_0 and A_1 are as obtained above.

4 Conclusion

In this chapter we constructed exact solutions of the Klein–Gordon–Zakharov equations via two different methods; the travelling wave approach and the simplest equation method. Firstly, we transformed the system of partial differential equations (PDEs) to a system of ordinary differential equations (ODEs) which we solved to obtain exact solutions of the KGZ equations given by (1) using three different approaches. The solutions obtained were travelling wave solutions.

Acknowledgments IEM and CMK would like to thank the Organizing Committee of 'International Conference: AMMCS-2013', Waterloo, Canada for their kind hospitality during the conference. IEM also thanks the Faculty Research Committee of FAST, North-West University, Mafikeng Campus for financial support.

References

1. Wang, T., Chen, J., Zhang, L.: Conservative difference methods for the Klein–Gordon–Zakharov equations. J. Comput. Appl. Math. **205,** 430–452 (2007)
2. Shi, Q., Xiao, Q., Liu, X.: Extended wave solutions for a nonlinear Klein–Gordon–Zakharov system. Appl. Math. Comput. **218,** 9922–9929 (2012)
3. Guo, B.L., Yuan, G.W.: Global smooth solution for the Klein–Gordon–Zakharov equations. J. Math. Phys. **36** (8), 4119–4124 (1995)
4. Ozawa, T., Tsutaya, K., Tsutsumi, Y.: Normal form and global solutions for the Klein–Gordon–Zakharov equations. Ann. Inst. H. Poincaré Anal. Non Linéaire. **12** (4), 459–503 (1995)
5. Tsutaya, K.: Global existence of small amplitude solutions for the Klein–Gordon–Zakharov equations. Nonlinear Anal. TMA. **27** (12), 1373–1380 (1996)

6. Adomian, G.: Non-perturbative solution of the Klein–Gordon–Zakharov equation. Appl. Math. Comput. **81** (1), 89–92 (1997)
7. Shang, Y., Huang, Y., Yuan, W.: New exact traveling wave solutions for the Klein–Gordon–Zakharov equations. Comput. Math. Appl. **56,** 1441–1450 (2008)
8. Chen, L.: Orbital stability of solitary waves for the Klein–Gordon–Zakharov equations. Acta. Math. Appl. Sin. (English Ser.) **15** (1), 54–64 (1999)
9. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products, 7th edn. Academic, New York, (2007)
10. Kudryashov, N.A.: Simplest equation method to look for exact solutions of nonlinear differential equations. Chaos Soliton Fract. **24,** 1217–1231 (2005)

Collision Effects of Solitary Waves for the Gardner Equation

Abdus Sattar Mia

Abstract We study the physical and collision properties of the combined KdV–mKdV solitons given by the Gardner equation which possess solitary wave solution characterized by *sech* function. A collision of the two solitary waves produces 2-soliton solution. We make a physical form of the 2-soliton solution where the fast soliton moves with speed c_1 and the slow soliton moves with speed c_2. In the collision described by the 2-soliton solution, the solitary waves preserve their shapes and speeds, but get a shift in position where the fast soliton overtakes the slow soliton if their speeds have same direction, and two solitons cross head-on if their speeds have opposite direction. For a collision there exist three different types of interactions which depend on the relative ratio c_1/c_2 of speeds and the relative orientation of the two solitary waves.

1 Introduction

The Gardner Eq. [5] also known as combined KdV–mKdV equation is given by

$$u_t + 2auu_x + 3bu^2u_x + u_{xxx} = 0 \tag{1}$$

where a, b are any arbitrary real constants and $u(x, t)$ is the amplitude of the ocean waves in shallow seas. The Eq. (1) is completely integrable [5, 10] with a Lax pair and inverse scattering transformation [1], and admits solitary-wave solution [4, 9, 10]. Wazwaz has derived multi-soliton solutions of the Gardner equation in his book [10] in an exponential form. In the 2-soliton solution, the solitary waves undergo a collision. We found that three types of collision exist: (I) a fast right-moving soliton with speed $c_1 > 0$ overtakes a slow right-moving soliton with speed $c_2 > 0$, (II) a fast left-moving soliton with speed $c_1 \in (-3/2, 0)$ overtakes a slow left-moving soliton $c_2 \in (-3/2, 0)$, and (III) a right-moving soliton with speed $c_1 > 0$ and a left-moving soliton $c_2 \in (-3/2, 0)$ collide head-on. We carry out an asymptotic analysis of the

A. S. Mia (✉)
University of Saskatchewan, Saskatoon, SK, Canada
e-mail: sattar_ju@yahoo.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_44

2-soliton solution to analyze the collision. For types (I) and (III), we find that the net effect of the collision is to produce a respective forward and backward shift in the positions of the fast and slow waves while for type (II) these shifts interchange the direction. In particular, these shifts are found to depend only on the speeds of the two waves.

We express the 2-soliton solution in the physical form that exhibits different interactions of two solitary waves during collision. Interactions have been classified in three categories depending on the speeds ratio and relative orientations of the solitary waves.

2 Gardner Solitary-Wave Solution

With the suitable values of the coefficients ($a = 3, b = 2$), the Gardner equation (1) becomes

$$u_t + 6uu_x + 6u^2 u_x + u_{xxx} = 0 \tag{2}$$

Using a transformation $u(x,t) = v(x,t) - 1/2$, the Eq. (2) can be written as

$$v_t - \frac{3}{2} v_x + 6v^2 v_x + v_{xxx} = 0 \tag{3}$$

which has the soliton solution of the rational form (see [3, 6–8])

$$v(x,t) = G/F \tag{4}$$

$$\text{with } \; G = 2(f_x g - g_x f), \quad F = f^2 + g^2$$

Then the 1-soliton solution for the Eq. (3) is given by ($f = e^{\theta},\; g = 1$)

$$v(x,t) = 2\kappa k e^{k\xi}/(1 + e^{2k\xi}), \quad \kappa = \pm 1 \tag{5}$$

where $\xi = x - ct$ is a travelling-wave coordinate centered at initial position $x = 0$ and moves with a speed $c = k^2 - 3/2$. So the 1-soliton solution for the Gardner equation:

$$u = -1/2 + \kappa\lambda \operatorname{sech}(\lambda(x - ct)) \tag{6}$$

with $\lambda = k = \sqrt{c + 3/2}$. For $k \in \mathbb{R}$, we see that $c \geq -3/2$.

This solution describes a stable travelling-wave that is single-spiked with up (or down) faced orientation for $\kappa = 1$ ($\kappa = -1$). Its height relative to $u = -1/2$ is proportional to $\pm\sqrt{c + 3/2}$, and its width is proportional to $\sqrt{2}/\sqrt{2c + 3}$. The first four conserved quantities of (2) are given by

$$\mathcal{M} = \int_{-\infty}^{\infty} u \, \mathrm{d}x, \quad \mathcal{P} = \int_{-\infty}^{\infty} u^2 \, \mathrm{d}x, \quad \mathcal{E} = \frac{1}{4} \int_{-\infty}^{\infty} (u^4 + 2u^3 - u_x^2) \, \mathrm{d}x \tag{7}$$

$$\mathcal{C} = \int_{-\infty}^{\infty} x(u + u^2) - t(3u^4 + 6u^3 + 3u^2 + 3uu_{xx}) \, \mathrm{d}x \tag{8}$$

first three of which are analogous to the KdV mass, momentum, and energy.

3 Gardner 2-Soliton Solution

The Hirota ansatz [10] for the 2-soliton solution of (3) are given by

$$f = e^{\theta_1} + e^{\theta_2}, \quad g = 1 - Ae^{\theta_1+\theta_2} \tag{9}$$

in terms of $\theta_1 = k_1\xi_1$ and $\theta_2 = k_2\xi_2$, $k_1, k_2 \in \mathbb{R}$, where the travelling-wave coordinates are given by $\xi_1 = x - c_1 t$, $\xi_2 = x - c_2 t$ with speeds $c_1 = k_1^2 - 3/2$, $c_2 = k_2^2 - 3/2$, and where A is given by (see [8, 10])

$$A = (k_1 - k_2)^2/(k_1 + k_2)^2 \tag{10}$$

The 2-soliton solution of (3) can be written in the rational form (4) in terms of

$$G = 2(\kappa_1\lambda_1 \exp(\lambda_1\xi_1)(1 + A\exp(2\lambda_2\xi_2)) + \kappa_2\lambda_2 \exp(\lambda_2\xi_2)(1 + A\exp(2\lambda_1\xi_1))) \tag{11}$$

$$\begin{aligned} F = {} & 1 + 2\kappa_1\kappa_2(1 - A)\exp(\lambda_1\xi_1 + \lambda_2\xi_2) + \exp(2\lambda_1\xi_1) + \exp(2\lambda_2\xi_2) \\ & + A^2 \exp(2(\lambda_1\xi_1 + \lambda_2\xi_2)) \end{aligned} \tag{12}$$

with $\lambda_1 = k_1 = \sqrt{c_1 + 3/2}$, $\lambda_2 = k_2 = \sqrt{c_2 + 3/2}$, $\kappa_1 = \pm 1$, $\kappa_2 = \pm 1$, and where $\xi_1 = x - c_1 t - a_1$, $\xi_2 = x - c_2 t - a_2$ are fast and slow travelling-wave coordinates having initial positions at $x = a_1$ and $x = a_2$ respectively.

3.1 Asymptotic Analysis

We study the asymptotic analysis of the 2-soliton solution (4, 11, 12) for both positive and negative directions ($t \to \pm\infty$).

For Positive Speeds and for Opposite Speeds ($c_1 > 0, c_2 \in (-3/2, 0)$). We keep the fast coordinate ξ_1 fixed and take $\xi_2 = \xi_1 + \eta$ in terms of ξ_1 and $\eta = t\Delta c - \Delta a$, where

$$\Delta c = c_1 - c_2 > 0 \tag{13}$$

is the resultant speed and $\Delta a = a_2 - a_1$ is the separation of the initial positions of the travelling-wave coordinates. Thus $t \to \pm\infty$ implies that $\eta \to \pm\infty$. The asymptotic expansion of F and G, by neglecting the dominated terms, yields

$$G \simeq \begin{cases} 2\kappa_1\lambda_1 A \exp(\lambda_1\xi_1)\exp(2\lambda_2\xi_1)\exp(2\lambda_2\eta) & \text{when } \eta \to +\infty \\ 2\kappa_1\lambda_1 \exp(\lambda_1\xi_1) & \text{when } \eta \to -\infty \end{cases}$$

$$F \simeq \begin{cases} \exp(2\lambda_2\xi_1)\exp(2\lambda_2\eta)(1 + A^2\exp(2\lambda_1\xi_1)) & \text{when } \eta \to +\infty \\ 1 + \exp(2\lambda_1\xi_1) & \text{when } \eta \to -\infty \end{cases}$$

Hence, by (4), $v \simeq \begin{cases} 2\kappa_1\lambda_1 A \exp(\lambda_1\xi_1)/(1 + A^2\exp(2\lambda_1\xi_1)), & \eta \to +\infty \\ 2\kappa_1\lambda_1 \exp(\lambda_1\xi_1)/(1 + \exp(2\lambda_1\xi_1)), & \eta \to -\infty \end{cases}$

Thus asymptotic future ($t \to +\infty$) and past ($t \to -\infty$) solutions are of the form of 1-soliton solution (5) with travelling-wave coordinates $\xi_1^{\pm} = \xi_1 - b_1^{\pm}$, where $b_1^{\pm}$, the position shifts of the waves, are given by

$$b_1^+ = -\ln(|A|)/\lambda_1 \text{ and } b_1^- = 0. \tag{14}$$

For an expansion in terms of ξ_2, we keep the slow coordinate ξ_2 be fixed and take $\xi_1 = \xi_2 - \eta$ where $\eta = t\Delta c - \Delta a$. Thus, $t \to \pm\infty$ implies that $\eta \to \pm\infty$. The asymptotic expansion of F and G, by neglecting the dominated terms, yields

$$G \simeq \begin{cases} 2\kappa_2\lambda_2 \exp(\lambda_2\xi_2) & \text{when } \eta \to +\infty \\ 2\kappa_2\lambda_2 A \exp(\lambda_2\xi_2) \exp(2\lambda_1\xi_2) \exp(-2\lambda_1\eta) & \text{when } \eta \to -\infty \end{cases}$$

$$F \simeq \begin{cases} 1 + \exp(2\lambda_2\xi_2) & \text{when } \eta \to +\infty \\ \exp(2\lambda_1\xi_2) \exp(-2\lambda_1\eta)(1 + A^2 \exp(2\lambda_2\xi_2)) & \text{when } \eta \to -\infty \end{cases}$$

Hence, by (4), $v \simeq \begin{cases} 2\kappa_2\lambda_2 \exp(\lambda_2\xi_2)/(1 + \exp(2\lambda_2\xi_2)), & \eta \to +\infty \\ 2\kappa_2\lambda_2 A \exp(\lambda_2\xi_2)/(1 + A^2 \exp(2\lambda_2\xi_2)), & \eta \to -\infty \end{cases}$

Thus, asymptotic future and past solutions are of the form of 1-soliton solution (5) with travelling-wave coordinates $\xi_2^{\pm} = \xi_2 - b_2^{\pm}$, where $b_2^{\pm}$, the position shifts of the waves, are given by

$$b_2^+ = 0 \text{ and } b_2^- = -\ln(|A|)/\lambda_2. \tag{15}$$

For Negative Speeds ($c_1 < c_2 < 0$). Since $\Delta c < 0$, we see that $t \to \pm\infty$ implies that $\eta \to \mp\infty$. The analysis require similar expansion as we did in previous case, but the asymptotic results are getting interchange with future ($t \to +\infty$) and past ($t \to -\infty$). Thus, the asymptotic future and past solutions are of the form of 1-soliton solution (5) with travelling-wave coordinates $\xi_1^{\pm} = \xi_1 - d_1^{\pm}$, where

$$d_1^+ = 0 \text{ and } d_1^- = -\ln(|A|)/\lambda_1, \quad \text{the position shifts of the waves} \tag{16}$$

In the expansion of ξ_2, the asymptotic future and past solutions are of the form of 1-soliton solution (5) with travelling-wave coordinates $\xi_2^{\pm} = \xi_2 - d_2^{\pm}$, where

$$d_2^+ = -\ln(|A|)/\lambda_2 \text{ and } d_2^- = 0. \tag{17}$$

3.1.1 Asymptotic Results

From the above asymptotic analysis we see that the solitary waves, for all cases, preserve their shapes and speeds, aside from getting a shift in position given by the formula in Theorem 3.1.1 and 3.1.2.

Theorem 3.1.1 *For an overtake collision with positive speeds, $c_1 > c_2 > 0$, or for a head-on collision with $c_1 > 0, c_2 \in (-3/2, 0)$, the fast and slow solitary waves* *undergo a respective forward and backward shift in position given by*

$$\Delta x_i = b_i^+ - b_i^- = (-1)^i \frac{2\sqrt{2}}{\sqrt{2c_i+3}} \ln\left(\frac{\sqrt{2c_1+3}-\sqrt{2c_2+3}}{\sqrt{2c_1+3}+\sqrt{2c_2+3}}\right), \quad i = 1, 2 \tag{18}$$

Theorem 3.1.2 *For an overtake collision with positive speeds,* $c_1 < c_2 < 0$, *the fast and slow solitary waves undergo a respective backward and forward shift in position given by*

$$\Delta x_i = d_i^+ - d_i^- = (-1)^{i+1} \frac{2\sqrt{2}}{\sqrt{2c_i+3}} \ln\left(\frac{\sqrt{2c_1+3}-\sqrt{2c_2+3}}{\sqrt{2c_1+3}+\sqrt{2c_2+3}}\right), \quad i = 1, 2 \tag{19}$$

4 Physical Solution and Interaction Properties

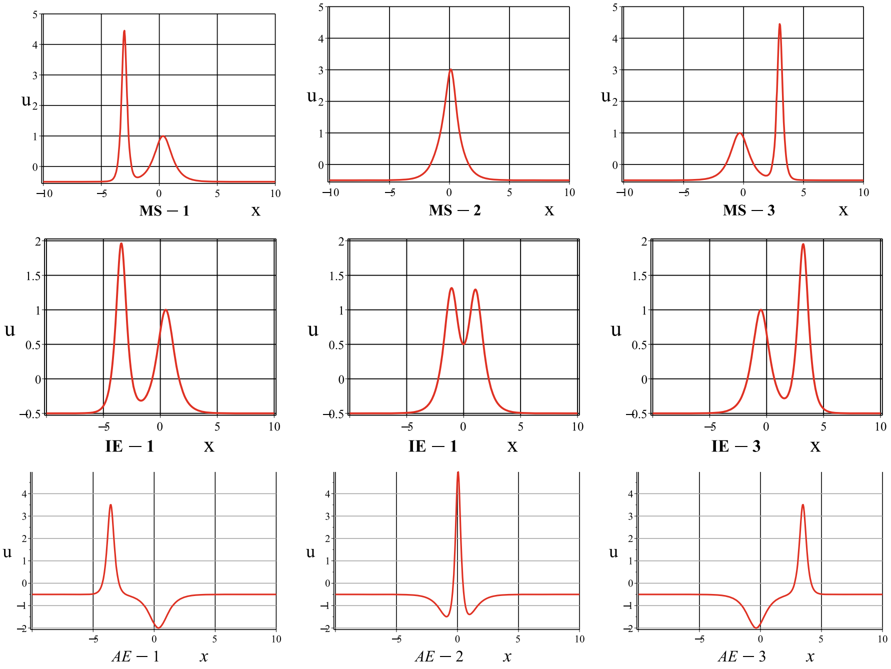
We use suitable time and space translations $t \to \tilde{t} - \epsilon, x \to \tilde{x} - \epsilon$ to shift the centers of the moving coordinates ξ_1, ξ_2 to the initial positions $a_1 = 0, a_2 = 0$, and then rewrite the 2-soliton solution ($u = -\frac{1}{2} + G/F$) in the physical form as

$$u(x,t) := -\frac{1}{2} + \frac{\sqrt{2}\alpha(\kappa_1\sqrt{2c_1+3}\cosh(\theta_2) + \kappa_2\sqrt{2c_2+3}\cosh(\theta_1))}{\kappa_1\kappa_2(1-\alpha^2) + \alpha^2\cosh(\theta_1+\theta_2) + \cosh(\theta_1-\theta_2)} \tag{20}$$

in terms of $\theta_1 = \sqrt{c_1 + 3/2}\,(x - c_1 t)$ and $\theta_2 = \sqrt{c_2 + 3/2}\,(x - c_2 t)$ where, $\alpha = (\sqrt{2c_1+3} - \sqrt{2c_2+3})/(\sqrt{2c_1+3} + \sqrt{2c_2+3})$. The physical solution (20) is invariant under a combined space and time reflection $x \to -x, t \to -t$. Since $u(-x) = u(x)$ when $t = 0$, the fast and slow soliton will have maximum interaction at $t = 0$. There are two cases to discuss.

(1) For an overtake collision: In the case of positive speeds, the 2-soliton solution $u(x, 0)$ will have either a single spike at x = 0 if $c_1/c_2 > 18$, a critical value, or a double spike about $x = 0$ if $c_1/c_2 < 18$. For a single spike, the fast and slow solitary waves interact by first merging together at $x = t = 0$ and then splitting apart, while in the case of a double spike, the fast and slow solitons interact by interchanging shapes and speeds at $x = t = 0$. The first one is called a merge-split [2] and we call the second one an inward–exchange interaction. These interactions are seen in the figures MS-1 to MS-3 and IE-1 to IE-3 respectively. For opposite orientations, the interaction shows that the slow solitary wave being first devoured continuously by the contacted end of the fast wave and then emitted from the other end of the fast wave. This is called an absorb–emit interaction [2]. See figures from AE-1 to AE-3.

(2) For a head-on collision: The 2-soliton solution $u(x, 0)$ has either a single spike at x = 0 if $|c_1/c_2| > 10$, or double spikes about $x = 0$ if $|c_1/c_2| < 10$. In this case, the interaction is the merge–split type at $x = t = 0$ for $|c_1/c_2| > 10$ while the interaction is the inward–exchange type at $x = t = 0$ for $|c_1/c_2| < 10$. For an



opposite orientation, the 2-soliton solution $u(x, 0)$ has an up- or down-faced spike at $x = 0$ with an exponentially diminishing end. The interaction between the fast and slow solitons, in this case, is the absorb–emit type at $x = t = 0$ for any speed ratio.

References

1. Ablowitz, M.J., Clarkson, P.A.: Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge (1991)
2. Anco, S.: http://lie.math.brocku.ca/sanco/solitons/mkdv
3. Drazin, P.G., Johnson, R.S.: Solitons: An Introduction. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (1989)
4. Fu, Z., Liu, S., Liu, S.: New kinds of solutions to Gardner equation. Chaos, Solitons & Fractals **20**(2), 301–309 (2004)
5. Slyunaev, A.V., Pelinovski, E.N.: Dynamics of large-amplitude solitons. J. Exp. Theor. Phys. **89**(1), 173–181 (1999)
6. Hietarinta, J.: A search for bilinear equations passing Hirota’s three-soliton condition. II. mKdV-type bilinear equations. J. Math. Phys. **28**(9), 2094–2101 (1987)
7. Hirota, R.: Exact solutions of the Korteweg-de Vries equation for multiple collisions of solitons. Phys. Rev. Lett. **27**(18), 1192–1194 (1971)
8. Hirota, R.: The Direct Method in Soliton Theory. Cambridge University Press, Cambridge (2004)
9. Wazwaz, A.M.: New solitons and kink solutions for the Gardner equation. Commun. Nonlinear Sci. Numer. Simul. **12**(8), 1395–1404 (2007)
10. Wazwaz, A.M.: Partial Differential Equations and Solitary Waves Theory. Higher Education Press, Beijing (2009)

Conservation Laws for a Generalized Coupled Boussinesq System of KdV–KdV Type

Tshepo Edward Mogorosi, Ben Muatjetjeja and Chaudry Masood Khalique

Abstract In this chapter, we consider a generalized coupled Boussinesq system of KdV–KdV type, which belongs to the class of Boussinesq systems modeling two-way propagation of long waves of small amplitude on the surface of an ideal fluid. We obtain conservation laws for this system using Noether theorem. Since this system does not have a Lagrangian, we increase the order of the partial differential equations by using the transformations $u = U_x$, $v = V_x$ and convert the Boussinesq system to a fourth-order system in U, V variables, which has a Lagrangian. Consequently, we find infinitely many nonlocal conserved quantities for our original Boussinesq system of KdV–KdV type.

1 Introduction

We consider the generalized coupled Boussinesq system of KdV–KdV type [11],

$$\begin{aligned} u_t + v_x + u_x v + u v_x + a v_{xxx} &= 0, \\ v_t + u_x + v v_x + c u_{xxx} &= 0, \end{aligned} \tag{1}$$

where a and c are arbitrary constants. The space and time variables x and t represent the position and the elapsed time, respectively along the channel, where $v(x,t)$ is the deviation of the free surface from its rest position and $u(x,t)$ is the horizontal velocity. System (1) is known to be a valid approximation of the full, two-dimensional Euler

T. E. Mogorosi (✉) · B. Muatjetjeja · C. M. Khalique
Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mafikeng Campus, Private Bag X 2046, Mmabatho 2735, Republic of South Africa
e-mail: tshepomogorosi@rocketmail.com

B. Muatjetjeja
e-mail: Ben.Muatjetjeja@nwu.ac.za

C. M. Khalique
e-mail: Masood.Khalique@nwu.ac.za

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_45

equations for fluid motion under the influence of gravity in suitably small amplitude, long wavelength regimes [1]. This system (1) falls under the family of Boussinesq systems derived in [2] and reduces to a symmetric hyperbolic system when the dispersive terms are dropped. The numerical solutions for Boussinesq systems have been investigated in [3, 4] using the standard Galerkin-finite element method.

In this chapter, we derive conservation laws for system (1). It is well-known that the conservation laws play a vital role in the study of nonlinear partial differential equations (PDEs). For variational problems the conservation laws can be constructed by means of Noether theorem [9]. Of course, the application of Noether theorem depends upon the existence of a Lagrangian. However, there are methods such as the Laplace direct method [7] and characteristics method [12] to obtain conservation laws for partial differential equations that do not have a Lagrangian. See also [5, 6, 10, 13]. Our system (1) does not have a Lagrangian. To use Noether theorem we increase the order of the system (1) such that it has a Lagrangian [8].

2 Conservation Laws for the Boussinesq System

Consider the generalized coupled Boussinesq system of KdV–KdV type

$$\begin{aligned} u_t + v_x + u_x v + u v_x + a v_{xxx} &= 0, \\ v_t + u_x + v v_x + c u_{xxx} &= 0, \end{aligned} \tag{2}$$

where a and c are arbitrary constants. This system does not have a Lagrangian. In order to apply Noether theorem we increase the order of this system by using the transformations $u = U_x$, $v = V_x$. Then the system (2) transforms to

$$\begin{aligned} U_{tx} + V_{xx} + U_{xx} V_x + U_x V_{xx} + a V_{xxxx} &= 0, \\ V_{xt} + U_{xx} + V_x V_{xx} + c U_{xxxx} &= 0 \end{aligned} \tag{3}$$

and has a Lagrangian. Since L given by

$$L = \frac{1}{2}\left\{ c U_{xx}^2 + a V_{xx}^2 - U_x^2 - V_x^2 - U_x V_x^2 - U_t V_x - U_x V_t \right\} \tag{4}$$

satisfies the Euler–Lagrange equations

$$\frac{\delta L}{\delta U} = 0 \quad \text{and} \quad \frac{\delta L}{\delta V} = 0, \tag{5}$$

we infer that L is a second-order Lagrangian for (3). Here $\delta/\delta U$ and $\delta/\delta V$ are the Euler–Lagrange operators defined by

$$\frac{\delta}{\delta U} = \frac{\partial}{\partial U} - D_t \frac{\partial}{\partial U_t} - D_x \frac{\partial}{\partial U_x} + D_t^2 \frac{\partial}{\partial U_{texttt}} + D_x^2 \frac{\partial}{\partial U_{xx}} + D_x D_t \frac{\partial}{\partial U_{tx}} - \cdots \tag{6}$$

and

$$\frac{\delta}{\delta V}=\frac{\partial}{\partial V}-D_t\frac{\partial}{\partial V_t}-D_x\frac{\partial}{\partial V_x}+D_t^2\frac{\partial}{\partial V_{texttt}}+D_x^2\frac{\partial}{\partial V_{xx}}+D_xD_t\frac{\partial}{\partial V_{tx}}-\cdots. \tag{7}$$

We now find Noether point symmetries for system (3) corresponding to L. Recall that the vector field

$$X=\xi^1(t,x,U,V)\frac{\partial}{\partial t}+\xi^2(t,x,U,V)\frac{\partial}{\partial x}+\eta^1(t,x,U,V)\frac{\partial}{\partial U}+\eta^2(t,x,U,V)\frac{\partial}{\partial V} \tag{8}$$

is a Noether point symmetry corresponding to the Lagrangian L if there exists gauge functions $B^1(t,x,U,V)$ and $B^2(t,x,U,V)$ such that

$$X^{[2]}(L)+\{D_t(\xi^1)+D_x(\xi^2)\}L=D_t(B^1)+D_x(B^2). \tag{9}$$

Here $X^{[2]}$ denotes the second prolongation of X and is defined as

$$\begin{aligned}X^{[2]}=&\,\xi^1(t,x,U,V)\frac{\partial}{\partial t}+\xi^2(t,x,U,V)\frac{\partial}{\partial x}+\eta^1(t,x,U,V)\frac{\partial}{\partial U}\\&+\eta^2(t,x,U,V)\frac{\partial}{\partial V}+\zeta_t^1\frac{\partial}{\partial U_t}+\zeta_t^2\frac{\partial}{\partial V_t}+\zeta_x^1\frac{\partial}{\partial U_x}+\zeta_x^2\frac{\partial}{\partial V_x}+\cdots,\end{aligned} \tag{10}$$

where

$$\begin{aligned}\zeta_t^1&=D_t(\eta^1)-U_tD_t(\xi^1)-U_xD_t(\xi^2), &\zeta_x^1&=D_x(\eta^1)-U_tD_x(\xi^1)-U_xD_x(\xi^2),\\ \zeta_t^2&=D_t(\eta^2)-V_tD_t(\xi^1)-V_xD_t(\xi^2), &\zeta_x^2&=D_x(\eta^2)-V_tD_x(\xi^1)-V_xD_x(\xi^2)\end{aligned}$$

and

$$\begin{aligned}D_t&=\frac{\partial}{\partial t}+U_t\frac{\partial}{\partial U}+V_t\frac{\partial}{\partial V}+U_{texttt}\frac{\partial}{\partial U_t}+V_{texttt}\frac{\partial}{\partial V_t}+U_{tx}\frac{\partial}{\partial U_x}+V_{tx}\frac{\partial}{\partial V_x}+\cdots,\\ D_x&=\frac{\partial}{\partial x}+U_x\frac{\partial}{\partial U}+V_x\frac{\partial}{\partial V}+U_{xx}\frac{\partial}{\partial U_x}+V_{xx}\frac{\partial}{\partial V_x}+U_{tx}\frac{\partial}{\partial U_t}+V_{tx}\frac{\partial}{\partial V_t}+\cdots.\end{aligned}$$

Inserting the value of L from (4) into Eq. (9) yields

$$\begin{aligned}&-\frac{1}{2}V_x\left[\eta_t^1+U_t\eta_U^1+V_t\eta_V^1-U_t\xi_t^1-U_t^2\xi_U^1-U_tV_t\xi_V^1-U_x\xi_t^2\right.\\&\left.-\,U_tU_t\xi_U^2-U_xV_t\xi_V^2\right]-\frac{1}{2}U_x\left[\eta_t^2+U_t\eta_U^2+V_t\eta_V^2-V_t\xi_t^1-U_tV_t\xi_U^1\right.\\&\left.-V_t^2\xi_V^1-V_x\xi_t^2-U_tV_x\xi_U^2-V_tV_x\xi_V^2\right]\\&-\left(U_x+\frac{1}{2}V_x^2+\frac{1}{2}V_t\right)\left[\eta_x^1+U_x\eta_U^1+V_x\eta_V^1\right.\\&\left.-U_t\xi_x^1-U_tU_x\xi_U^1-U_tV_x\xi_V^1-U_x\xi_x^2-U_x^2\xi_U^2-U_xV_x\xi_V^2\right]\end{aligned}$$

$$
\begin{aligned}
&-\left(V_x+\frac{1}{2}U_t+U_xV_x\right)\left[\eta^2_x+U_x\eta^2_U+V_x\eta^2_V-V_t\xi^1_x-U_xV_t\xi^1_U\right.\\
&-\left.V_tV_x\xi^1_V-V_x\xi^2_x-U_xV_x\xi^2_U-V^2_x\xi^2_V\right]+cU_{xx}\left[D^2_x\eta^1-U_tD^2_x\xi^1\right.\\
&-\left.U_xD^2_x\xi^2-2U_{tx}\left(\xi^1_x+U_x\xi^1_U+V_x\xi^1_V\right)-2U_{xx}\left(\xi^2_x+U_x\xi^2_U+V_x\xi^2_V\right)\right]\\
&+aV_{xx}[D^2_x\eta^2-V_tD^2_x\xi^1-V_xD^2_x\xi^2-2V_{tx}\left(\xi^1_x+U_x\xi^1_U+V_x\xi^1_V\right)\\
&-2V_{xx}\left(\xi^2_x+U_x\xi^2_U+V_x\xi^2_V\right)]\\
&+\frac{1}{2}\left[cU^2_{xx}+aV^2_{xx}-U^2_x-V^2_x-U_xV^2_x-U_tV_x-U_xV_t\right]\left[\xi^1_t+U_t\xi^1_U\right.\\
&+\left.V_t\xi^1_V+\xi^2_x+U_x\xi^2_U+V_x\xi^2_V\right]\\
&=B^1_t+U_tB^1_U+V_tB^1_V+B^2_x+U_xB^2_U+V_xB^2_V.
\end{aligned}
\tag{11}
$$

The splitting of (11) with respect to different combinations of derivatives of U and V results in an over-determined system of PDEs for ξ^1, ξ^2, η^1, η^2, B^1, and B^2. After some tedious calculations, the solution of the system yields the following Noether point symmetries and gauge terms:

$$
\begin{aligned}
\xi^1&=c_1,\\
\xi^2&=c_2,\\
\eta^1&=H(t),\\
\eta^2&=J(t),\\
B^1&=w(t,x),\\
B^2&=-\frac{1}{2}UJ'(t)-\frac{1}{2}VH'(t)+z(t,x),\\
w_t+z_x&=0.
\end{aligned}
\tag{12}
$$

We can set $w=0$, $z=0$ as they contribute to the trivial part of the conserved vector.

Recall that the formulae for the conserved vector (T^1,T^2) for the second-order Lagrangian L [8, 9] are given by

$$
\begin{aligned}
T^1=-B^1+\xi^1L&+W^1\left[\frac{\partial L}{\partial U_t}-D_t\frac{\partial L}{\partial U_{textttt}}-D_x\frac{\partial L}{\partial U_{tx}}\ldots,\right]\\
&+W^2\left[\frac{\partial L}{\partial V_t}-D_t\frac{\partial L}{\partial V_{xt}}-D_x\frac{\partial L}{\partial V_{textttt}}\ldots,\right]\\
&+D_t(W^1)\frac{\partial L}{\partial U_{textttt}}+D_t(W^2)\frac{\partial L}{\partial V_{textttt}},
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
T^2 = -B^2 + \xi^2 L + W^1 \left[\frac{\partial L}{\partial U_x} - D_t \frac{\partial L}{\partial U_{xt}} - D_x \frac{\partial L}{\partial U_{xx}} \dots, \right] \\
+ W^2 \left[\frac{\partial L}{\partial V_x} - D_t \frac{\partial L}{\partial V_{xt}} - D_x \frac{\partial L}{\partial V_{xx}} \dots, \right] \\
+ D_x(W^1) \frac{\partial L}{\partial U_{xx}} + D_x(W^2) \frac{\partial L}{\partial V_{xx}},
\end{aligned}
\tag{14}
$$

where $W^1 = \eta^1 - U_t \xi^1 - U_x \xi^2$ and $W^2 = \eta^2 - V_t \xi^1 - V_x \xi^2$ are the Lie-characterictic functions.

Thus, Eqs. (13) and (14) together with (12) and $u = U_x$, $v = V_x$ yield the following independent conserved vectors for system (2):

$$
\begin{aligned}
T_1^1 &= \frac{cu_x^2}{2} + \frac{av_x^2}{2} - \frac{u^2}{2} - \frac{v^2}{2} - \frac{uv^2}{2}, \\
T_1^2 &= u \int u_t dx + v \int v_t dx + uv \int v_t dx + \frac{v^2}{2} \int u_t dx + \int u_t dx \int v_t dx \\
&\quad + cu_{xx} \int u_t dx + av_{xx} \int v_t dx - cu_t u_x - av_t v_x;
\end{aligned}
\tag{15}
$$

$$
\begin{aligned}
T_2^1 &= uv, \\
T_2^2 &= cuu_{xx} + avv_{xx} - \frac{cu_x^2}{2} - \frac{av_x^2}{2} + uv^2 + \frac{u^2}{2} + \frac{v^2}{2}
\end{aligned}
\tag{16}
$$

and for the arbitrary functions $H(t)$ and $J(t)$

$$
\begin{aligned}
T_{(E,F)}^1 &= -\frac{v}{2} H(t) - \frac{u}{2} J(t), \\
T_{(E,F)}^2 &= -H(t) \left[u + \frac{v^2}{2} + cu_{xx} \right] - J(t) \left[v + uv + av_{xx} \right] - \frac{1}{2} J(t) \int u_t dx \\
&\quad - \frac{1}{2} H(t) \int v_t dx + \frac{1}{2} J'(t) \int u dx + \frac{1}{2} H'(t) \int v dx.
\end{aligned}
\tag{17}
$$

Notice that the conserved vector (16) is a local conserved vector and (15) is a nonlocal conserved vector for the system (2). We can obtain two special cases from the conserved vector (17) by letting $H(t) = 1$ and $J(t) = 0$, which gives a nonlocal conserved vector

$$
\begin{aligned}
T_3^1 &= -\frac{v}{2}, \\
T_3^2 &= -u - \frac{v^2}{2} - cu_{xx} - \frac{1}{2} \int v_t dx,
\end{aligned}
$$

and for $H(t) = 0$ and $J(t) = 1$, which also gives a nonlocal conserved vector

$$T_4^1 = -\frac{u}{2},$$
$$T_4^2 = -v - uv - av_{xx} - \frac{1}{2}\int u_t dx.$$

We note that since the functions $H(t)$ and $J(t)$ are arbitrary, one obtains infinitely many nonlocal conservation laws for the system (2).

3 Conclusion

In this chapter, we studied the third-order generalized coupled Boussinesq system of KdV–KdV type. In order to apply Noether theorem, the transformations $u = U_x$, $v = V_x$ were utilized. The system was transformed to the fourth-order system in U, V variables, which admitted a Lagrangian. Noether's approach was then used to derive the conservation laws in U, V variables. Finally, the inverse transformations $U = \int udx$, $V = \int vdx$ were used to obtain the conservation laws for the original coupled Boussinesq systems of KdV–KdV type. The conservation laws obtained consist of one local and infinite number of nonlocal conserved vectors.

Acknowledgement BM and CMK would like to thank the Organizing Committee of "International Conference: AMMCS-2013," Waterloo, Canada for their kind hospitality during the conference. BM also thanks the Faculty Research Committee of FAST, North-West University, and ETM thanks SANHARP for financial support. We thank the referee for the useful comments, which enhanced the presentation and results of the chapter.

References

1. Bona, J.L., Chen, M., Saut, J.C.: Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. I. Derivation and linear theory. J. Nonlinear Sci. **12,** 283–318 (2002)
2. Bona, J.L., Colin, T., Lannes, D.: Long wave approximations for water waves. Arch. Ration. Mech. Anal. **178,** 373–410 (2005)
3. Bona, J.L., Dougalis, V.A., Mitsotakis, D.E.: Numerical solution of Boussinesq systems of KdV–KdV type: II. Evolution of radiating solitary waves. Nonlinearity **21,** 1–24 (2008)
4. Bona, J.L., Dougalis, V.A., Mitsotakis, D.E.: Numerical solution of KdV–KdV systems of Boussinesq equations: I. The numerical scheme and existence of generalized solitary waves. Math. Comput. Simul. **74,** 214–228 (2007)
5. Kara, A.H., Mahomed, F.M.: Relationship between symmetries and conservation laws. Int. J. Theor. Phys. **39,** 23–40 (2000)
6. Kara, A.H., Mahomed, F.M.: Noether-type symmetries and conservation laws via partial Lagragians. Nonlinear Dyn. **45,** 367–383 (2006)
7. Laplace, P.S.: Traité de Mécanique Céleste, vol. 1, Paris (1798). (English transl., Celestial mechanics, New York, (1966))

8. Naz, R., Mahomed, F.M., Hayat, T.: Conservation laws for third-order variant Boussinesq system. Appl. Math. Lett. **23,** 883–886 (2010)
9. Noether, E.: Invariante variationsprobleme. Nachr. König. Gesell. Wiss. Göttingen Math.-phys. Kl. Heft **2**, 235–257 (1918). (The English traslation in transport Theory and Statistical Physics **1**, 186–207 (1971))
10. Olver, P.J.: Applications of Lie Groups to Differential Equations. Springer, New York (1993)
11. Pazoto, A.F., Rosier, L.: Stabilization of a Boussinesq system of KdV–KdV type. Syst. Control Lett. **57,** 595–601 (2008)
12. Steudel, H.: Uber die zuordnung zwischen invarianzeigenschaften und erhaltungssatzen. Z. Naturforsch. **A 17,** 129–132 (1962)
13. Wolf, T.: A comparison of four approaches to the calculation of conservation laws. Eur. J. Appl. Math. **13,** 129–152 (2002)

Exact Solutions of a Coupled Boussinesq Equation

Dimpho Millicent Mothibi and Chaudry Masood Khalique

Abstract In this chapter, (G'/G)-expansion method is employed to derive new exact solutions of a coupled Boussinesq equation. Three types of solutions are obtained, namely, hyperbolic function solutions, trigonometric function solutions and rational solutions. These solutions are travelling wave solutions.

1 Introduction

Many physical phenomena in science and engineering are modelled by nonlinear evolution equations. Also many nonlinear phenomena lead to coupled nonlinear evolution equations. In the past few decades various methods have been developed by scientists to find exact solutions of nonlinear evolution equations and coupled nonlinear evolution equations. These include the inverse scattering transform method [1], Bäcklund transformation [2], Darboux transformation [3], Hirota's bilinear method [4], the (G'/G)-expansion method [5], the reduction mKdV equation method [6], the sine–cosine method [7], the Jacobi elliptic function expansion method [8, 9], the F-expansion method [10], the exp-function expansion method [11] and the Lie symmetry method [12–15].

In this chapter, we study the coupled Boussinesq equation [16]

$$\begin{aligned} u_t + uu_x + v_x + au_{xxt} &= 0, \\ v_t + (uv)_x + bu_{xxx} &= 0, \end{aligned} \tag{1}$$

where u and v are real-valued scalar functions, t is time and x is a spatial variable and derive the travelling wave solutions of (1) by using the (G'/G)-expansion method.

D. M. Mothibi (✉) · C. M. Khalique
Department of Mathematical Sciences, International Institute for Symmetry Analysis and Mathematical Modelling, North-West University, Mafikeng Campus, Private Bag X 2046, Mmabatho 2735, Republic of South Africa
e-mail: Dimpho.Mothibi@nwu.ac.za,

C. M. Khalique
e-mail: Masood.Khalique@nwu.ac.za

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_46

2 Exact Solutions of a Coupled Boussinesq Equation

In this section we employ the (G'/G)-expansion method and construct the travelling wave solutions of the coupled Boussinesq equation (1).

As a first step we transform the coupled Boussinesq equations (1) to nonlinear ordinary differential equations (ODEs) using the travelling wave variable

$$u(t,x) = U(\xi),\ \ v(t,x) = V(\xi),\ \ \text{where}\ \ \xi = x - ct. \tag{2}$$

Using the above transformations, equations (1) transform to the nonlinear ODEs

$$\begin{aligned} acU''' + cU' - UU' - V' &= 0, \\ bU''' - cV' + VU' + UV' &= 0, \end{aligned} \tag{3}$$

where the primes denote the derivative with respect to ξ.

The (G'/G)-expansion method assumes the solutions of equations (3) to be of the form

$$U(\xi) = \sum_{i=0}^{M} \alpha_i (G'/G)^i \quad \text{and} \quad V(\xi) = \sum_{i=0}^{M} \beta_i (G'/G)^i, \tag{4}$$

where α_i, β_i, $i = 0, 1, \cdots, M$ are parameters to be determined and $G(\xi)$ satisfies the second-order linear ODE with constant coefficients, viz.,

$$G'' + \lambda G' + \mu G = 0, \tag{5}$$

where λ and μ are constants.

The balancing procedure yields $M = 2$, so the solutions of the ODEs (3) are of the form

$$\begin{aligned} U(\xi) &= \alpha_2 (G'/G)^2 + \alpha_1 (G'/G) + \alpha_0, \\ V(\xi) &= \beta_2 (G'/G)^2 + \beta_1 (G'/G) + \beta_0. \end{aligned} \tag{6}$$

Substituting (6) into (3) and making use of (5), and then collecting all terms with same powers of (G'/G) and equating each coefficient to zero, yields a system of algebraic equations. Solving this system of algebraic equations, using Mathematica, we obtain the following solutions:

$$\alpha_0 = \frac{2ac^2(a\lambda^2 + 8a\mu + 1) + b}{2ac},\ \alpha_1 = 12ac\lambda,\ \alpha_2 = 12ac,$$

$$\beta_0 = \frac{b^2 - 2a^2bc^2\lambda^2 - 16a^2bc^2\mu}{4a^2c^2},\ \beta_1 = -6b\lambda,\ \beta_2 = -6b.$$

Substituting the values of α's and β's and the corresponding solutions of ODE (5) into (6), we obtain the following three types of travelling wave solutions of equation (1):

Case 1: When $\lambda^2 - 4\mu > 0$, we obtain the hyperbolic function solutions

$$u_1(t,x) = \frac{2ac^2(a\lambda^2 + 8a\mu + 1) + b}{2ac}$$

$$+ 12ac\lambda\left[-\frac{\lambda}{2} + \delta_1\left(\frac{C_1 \sinh(\delta_1\xi) + C_2 \cosh(\delta_1\xi)}{C_1 \cosh(\delta_1\xi) + C_2 \sinh(\delta_1\xi)}\right)\right]$$

$$+ 12ac\left[-\frac{\lambda}{2} + \delta_1\left(\frac{C_1 \sinh(\delta_1\xi) + C_2 \cosh(\delta_1\xi)}{C_1 \cosh(\delta_1\xi) + C_2 \sinh(\delta_1\xi)}\right)\right]^2,$$

$$v_1(t,x) = \frac{b^2 - 2a^2bc^2\lambda^2 - 16a^2bc^2\mu}{4a^2c^2}$$

$$- 6b\lambda\left[-\frac{\lambda}{2} + \delta_1\left(\frac{C_1 \sinh(\delta_1\xi) + C_2 \cosh(\delta_1\xi)}{C_1 \cosh(\delta_1\xi) + C_2 \sinh(\delta_1\xi)}\right)\right],$$

$$- 6b\left[-\frac{\lambda}{2} + \delta_1\left(\frac{C_1 \sinh(\delta_1\xi) + C_2 \cosh(\delta_1\xi)}{C_1 \cosh(\delta_1\xi) + C_2 \sinh(\delta_1\xi)}\right)\right]^2,$$

where $\xi = x - ct$, $\delta_1 = \frac{1}{2}\sqrt{\lambda^2 - 4\mu}$, C_1 and C_2 are arbitrary constants.

Case 2: When $\lambda^2 - 4\mu < 0$, we obtain the trigonometric function solutions

$$u_2(t,x) = \frac{2ac^2(a\lambda^2 + 8a\mu + 1) + b}{2ac}$$

$$+ 12ac\lambda\left[-\frac{\lambda}{2} + \delta_2\frac{-C_1 \sin(\delta_2\xi) + C_2 \cos(\delta_2\xi)}{C_1 \cos(\delta_2\xi) + C_2 \sin(\delta_2\xi)}\right]$$

$$+ 12ac\left[-\frac{\lambda}{2} + \delta_2\frac{-C_1 \sin(\delta_2\xi) + C_2 \cos(\delta_2\xi)}{C_1 \cos(\delta_2\xi) + C_2 \sin(\delta_2\xi)}\right]^2,$$

$$v_2(t,x) = \frac{b^2 - 2a^2bc^2\lambda^2 - 16a^2bc^2\mu}{4a^2c^2}$$

$$- 6b\lambda\left[-\frac{\lambda}{2} + \delta_2\frac{-C_1 \sin(\delta_2\xi) + C_2 \cos(\delta_2\xi)}{C_1 \cos(\delta_2\xi) + C_2 \sin(\delta_2\xi)}\right]$$

$$- 6b\left[-\frac{\lambda}{2} + \delta_2\frac{-C_1 \sin(\delta_2\xi) + C_2 \cos(\delta_2\xi)}{C_1 \cos(\delta_2\xi) + C_2 \sin(\delta_2\xi)}\right]^2,$$

where $\xi = x - ct$, $\delta_2 = \frac{1}{2}\sqrt{4\mu - \lambda^2}$, C_1 and C_2 are arbitrary constants.

Case 3: When $\lambda^2 - 4\mu = 0$, we obtain the rational solutions

$$\begin{aligned} u_3(t,x) &= \frac{2ac^2(a\lambda^2 + 8a\mu + 1) + b}{2ac} + 12ac\lambda\left[-\frac{\lambda}{2} + \frac{C_2}{C_1 + C_2\xi}\right] \\ &+ 12ac\left[-\frac{\lambda}{2} + \frac{C_2}{C_1 + C_2\xi}\right]^2, \\ v_3(t,x) &= \frac{b^2 - 2a^2bc^2\lambda^2 - 16a^2bc^2\mu}{4a^2c^2} - 6b\lambda\left[-\frac{\lambda}{2} + \frac{C_2}{C_1 + C_2\xi}\right] \\ &- 6b\left[-\frac{\lambda}{2} + \frac{C_2}{C_1 + C_2\xi}\right]^2, \end{aligned}$$

where $\xi = x - ct$, C_1 and C_2 are arbitrary constants.

It should be noted that the solutions obtained in this chapter by (G'/G)-expansion method are more general than the solutions obtained in [16].

3 Conclusion

In this chapter, we analysed a coupled Boussinesq equation that appears in many scientific fields. The (G'/G)-expansion method was effectively used to derive exact travelling wave solutions of the coupled Boussinesq equation. The solutions obtained were expressed in the form of hyperbolic function, trigonometric function and rational solutions.

Acknowledgement DMM and CMK would like to thank the Organizing Committee of "International Conference: AMMCS-2013," Waterloo, Canada for their kind hospitality during the conference. DMM also thanks the Material Science Innovation and Modelling (MaSIM), a Research Focus Area in the Faculty of Agriculture, Science and Technology, North-West University, Mafikeng Campus for the financial support.

References

1. Ablowitz, M.J., Clarkson, P.A.: Soliton, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge (1991)
2. Gu, C.H.: Soliton Theory and Its Application. Zhejiang Science and Technology Press, Zhejiang (1990)
3. Matveev, V.B., Salle, M.A.: Darboux Transformation and Soliton. Springer-Verlag, Berlin (1991)
4. Hirota, R.: The Direct Method in Soliton Theory. Cambridge University Press, Cambridge (2004)
5. Wang, M., Xiangzheng, L.X., Jinliang, Z.J.: The (G'/G)-expansion method and travelling wave solutions of nonlinear evolution equations in mathematical physics. Phys. Lett. **A372**, 417–423 (2008)

6. Yan, Z.Y.: A reduction mKdV method with symbolic computation to construct new doubly-periodic solutions for nonlinear wave equations. Int. J. Mod. Phys. C. **14**, 661–672 (2003)
7. Wazwaz, M.: The tanh and sine–cosine method for compact and noncompact solutions of nonlinear Klein Gordon equation. Appl. Math. Comput. **167**, 1179–1195 (2005)
8. Lu, D.C.: Jacobi elliptic functions solutions for two variant Boussinesq equations. Chaos Soliton Fract. **24**, 1373–1385 (2005)
9. Yan, Z.Y.: Abundant families of Jacobi elliptic functions of the (2+1) dimensional integrable Davey-Stewartson-type equation via a new method. Chaos Soliton Fract. **18**, 299–309 (2003)
10. Wang, M., Li, X.: Extended F-expansion and periodic wave solutions for the generalized Zakharov equations. Phys. Lett. A. **343**, 48–54 (2005)
11. He, J.H., Wu, X.H.: Exp-function method for nonlinear wave equations. Chaos Soliton Fract. **30**, 70 (2006)
12. Bluman, G.W., Kumei, S.: Symmetries and Differential Equations (Applied Mathematical Sciences, vol. 81). Springer-Verlag, New York (1989)
13. Olver, P.J.: Applications of Lie Groups to Differential Equations (Graduate Texts in Mathematics) vol. 107, 2nd edn. Springer-Verlag, Berlin (1993)
14. Ibragimov, N.H.: CRC Handbook of Lie Group Analysis of Differential Equations, Vol 1–3. CRC Press, Boca Raton (1994–1996)
15. Ovsiannikov, L.V.: Group Analysis of Differential Equations. Academic, New York (1982). (English translation by W.F. Ames)
16. Wazwaz, A.M.: Solitons and periodic wave solutions for couples nonlinear equations. Int. J. Nonlinear Sci. **14**, 266–277 (2012)

Recent Advances in Error Control B-spline Gaussian Collocation Software for PDEs

Paul Muir and Jack Pew

Abstract In this chapter we briefly review recent advances in Error Control B-spline Gaussian Collocation software for the numerical solution of 1D parabolic partial differential equations (PDEs). BACOL and BACOLR, two packages of this type, developed over the last decade, have been shown to be efficient, reliable, and robust, especially for problems having solutions with sharp moving layers and for stringent tolerances. These packages use high order methods in time and space and feature *adaptive control of high order estimates of the temporal and spatial errors*. The spatial error estimates require the computation of a second collocation solution, which introduces a substantial computational overhead. In order to address this issue, a new software package, called BACOLI, has recently been developed (through a substantial modification of BACOL) in which the computation of the second collocation solution is replaced by the computation of a high order interpolant. Numerical results have shown that BACOLI computes spatial error estimates that are generally of comparable quality to those computed by BACOL and that the new code is generally substantially more efficient than BACOL.

1 Introduction

In this chapter we review recent work on new adaptive Error Control B-spline Gaussian Collocation software for the efficient numerical solution of systems of 1D parabolic PDEs. BACOL [11] and BACOLR [13], two packages of this type, developed over the last decade, use high order methods in time and space and feature *adaptive control of high order estimates of the temporal and spatial errors*. They have been shown to be efficient, reliable, and robust, especially for problems having solutions with sharp moving layers and for stringent tolerances [12]. These packages employ adaptive Error Control B-spline Gaussian Collocation for the spatial discretization of the partial differential equations (PDEs), leading to a system of time-dependent differential–algebraic equations (DAEs), which is solved in BACOL

P. Muir (✉) · J. Pew
Saint Mary's University, Halifax, NS, Canada
e-mail: muir@smu.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_47

using DASSL [5] and in BACOLR using RADAU5 [8]. Control of estimates of the temporal error is handled by the DAE solver. Control of spatial error estimates is handled using adaptive spatial mesh refinement based on high order estimates of the spatial error. These spatial error estimates are obtained by computing a second collocation solution (at substantial additional cost). Recent work to address this cost issue has focused on interpolation based schemes that allow a spatial error estimate to be obtained without the need for the computation of a second collocation solution. These schemes, called the superconvergent interpolant (SCI) scheme [1] and the lower order interpolant (LOI) scheme [3], have been implemented in the recently developed software package, BACOLI [10], obtained through a substantial modification of BACOL.

The problem class assumed by BACOL, BACOLR, and BACOLI has the form

$$\underline{u}_t(x,t) = \underline{f}(x,t,\underline{u}(x,t),\underline{u}_x(x,t),\underline{u}_{xx}(x,t)), \tag{1}$$

with initial and boundary conditions of the form

$$\underline{u}(x,t_0) = \underline{u}_0(x), \quad \underline{b}_L(t,\underline{u}(a,t),\underline{u}_x(a,t)), \quad \underline{b}_R(t,\underline{u}(b,t),\underline{u}_x(b,t)), \tag{2}$$

where $\underline{u}$, $\underline{f}$, $\underline{u}_0$, $\underline{b}_L$, and $\underline{b}_R$ are vector functions with *NPDE* components (where NPDE is the number of PDEs).

This chapter is organized as follows. In Sect. 2, we briefly review the adaptive Error Control B-spline Gaussian Collocation algorithm implemented in BACOL/BACOLR.

Section 3 briefly describes the SCI and LOI schemes employed in the new BACOLI code while Sect. 4 presents numerical results comparing the accuracy of the BACOL and BACOLI error estimates and the overall efficiency of the codes.

2 BACOL/BACOLR Error Control B-spline Gaussian Collocation

Assuming a spatial mesh of *NINT* subintervals that partitions $[a,b]$, the B-spline collocation algorithm employed by BACOL/BACOLR assumes that the collocation solution, $\underline{U}(x,t)$, is represented as a (vector) linear combination of known C^1-continuous piecewise polynomials of degree p on each spatial subinterval (represented in terms of a B-spline basis [6]) having the form

$$\underline{U}(x,t) = \sum_{i=1}^{NC} \underline{y}_i(t) B_i(x), \tag{3}$$

where $B_i(t)$ is the ith B-spline basis function, $\underline{y}_i(t)$ is the vector of corresponding B-spline coefficients, and $NC = NINT(p-1)+2$. The coefficients, $\underline{y}_i(t)$, are determined by requiring the collocation solution to satisfy the boundary conditions (at $x=a$ and $x=b$) and the PDE at $p-1$ collocation points per subinterval that are the images of the set of $p-1$ Gauss points [4] on [0, 1]. This gives the DAE system

$$\underline{0} = \underline{b}_L(a, t, \underline{U}(a, t), \underline{U}_x(a, t)), \tag{4}$$

$$\underline{U}_t(\xi_j, t) = \underline{f}(\xi_j, t, \underline{U}(\xi_j, t), \underline{U}_x(\xi_j, t), \underline{U}_{xx}(\xi_j, t)), \tag{5}$$

$$\underline{0} = \underline{b}_R(b, t, \underline{U}(b, t), \underline{U}_x(b, t)), \tag{6}$$

where ξ_j is the jth collocation point. As mentioned in the previous section, this DAE system is solved in BACOL using DASSL and in BACOLR using RADAU5. The spatial error estimate is obtained by using the same B-spline collocation algorithm, with a B-spline basis of degree $p + 1$, to obtain a second (higher order) collocation solution, $\bar{\underline{U}}(x, t)$. At the end of each timestep (let t be the current time), a norm of the difference between $\underline{U}(x, t)$ and $\bar{\underline{U}}(x, t)$ is computed to provide a spatial error estimate for $\underline{U}(x, t)$ over $[a, b]$. BACOL/BACOLR accepts $\underline{U}(x, t)$ provided that this spatial error satisfies the user tolerance. If it does not, the codes compute a second spatial error estimate, again based on the difference between the two collocation solutions, giving an estimate of the spatial error on each spatial subinterval, which is then used as the basis for a spatial remeshing, and the timestep is repeated. See [11] for further details.

As mentioned in the previous section, the computation of the second collocation solution represents a significant computational cost, essentially doubling the overall cost of the algorithm.

3 Review of the SCI/LOI Schemes and BACOLI

In order to improve the efficiency of the BACOL/BACOLR spatial error estimate, [1] and [3] consider, respectively, the SCI and LOI schemes, in which one of the two collocation solutions computed by BACOL/BACOLR is replaced with an interpolant.

The SCI scheme is based on theoretical results [7] that prove that, for (1), the collocation solution and its first spatial derivative are superconvergent at the spatial mesh points. Furthermore, based on similar theory for collocation methods for boundary value ordinary differential equations (ODEs)—see, e.g., [4]—and from experimental results for (1) [2], it is apparent that there are several (known) points internal to each subinterval, where the collocation solution is also superconvergent. The SCI scheme replaces the *higher* order collocation solution with a C^1-continuous piecewise polynomial, of the same order, specified on each subinterval by requiring it to interpolate the superconvergent meshpoint collocation solution and derivative values, the internal superconvergent collocation solution values, and the closest superconvergent collocation solution values from within each adjacent subinterval. Because the SCI interpolates at points from multiple subintervals, the interpolation error can be large when adjacent subinterval size ratios are large. See [1] for further details.

In contrast, the LOI scheme replaces the *lower* order collocation solution with a C^1-continuous piecewise polynomial specified on each subinterval by requiring it to interpolate the *higher* order collocation solution and its first spatial derivative at the mesh points and collocation solution at certain points within each subinterval such that the interpolation error of the resultant interpolant agrees asymptotically with the collocation error of the lower order collocation solution. See [9] for related work and [3] for further details.

As mentioned earlier, these schemes are implemented in the new code, BACOLI, in which only one collocation solution is computed and there is the option to obtain the spatial error estimate using either the SCI or LOI scheme. When BACOLI uses the SCI scheme it computes the lower order collocation solution and controls an error estimate for this solution; this is called standard (ST) error control mode. The original BACOL code also uses ST error control mode. When BACOLI uses the LOI scheme, it computes the higher order collocation solution but controls an error estimate for the lower order collocation solution; this is known as local extrapolation (LE) error control mode. If the original BACOL code were to be modified slightly to return the higher order collocation solution rather than the lower order collocation solution, it would be using LE error control mode.

4 Numerical Results

A standard test problem of the form (1) is the One Layer Burgers' Equation (OLBE):

$$u_t = \varepsilon u_{xx} - u u_x, \tag{7}$$

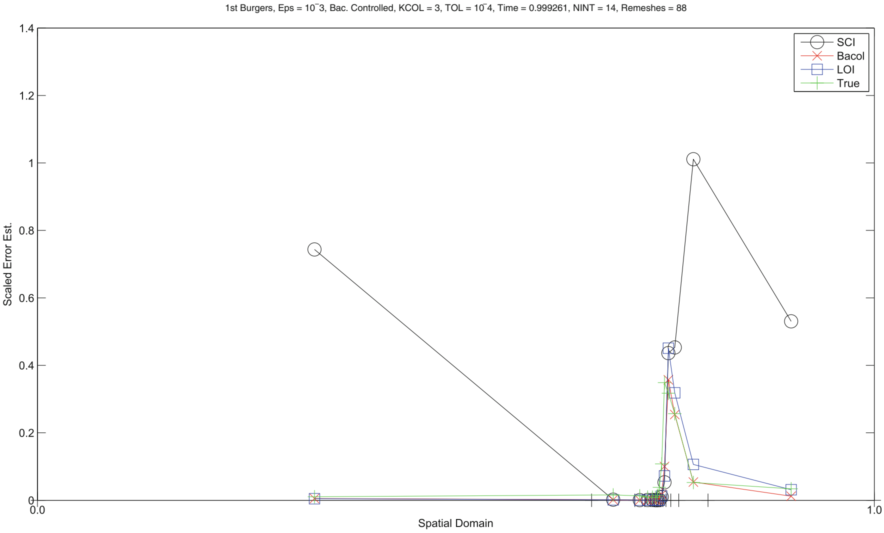


Fig. 1 BACOL, SCI, LOI error estimates and the true error for OLBE ($\varepsilon = 10^{-3}$) at $t = 1$, with $p = 4$, $tol = 10^{-4}$. (BACOL error estimates control the mesh.) The error estimates are in good agreement with each other and the true error except for the SCI estimates on subintervals for which the adjacent subinterval size ratios are large

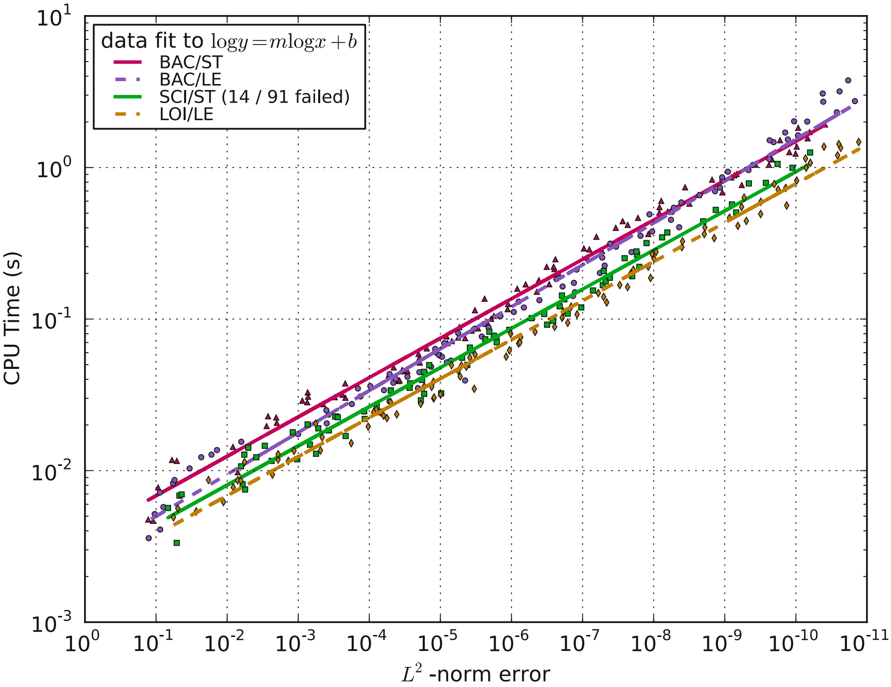


Fig. 2 Accuracy vs. time for OLBE ($\varepsilon = 10^{-3}$) at $t = 1$, for BACOL in ST and LE error control modes (BAC/ST, BAC/LE) and BACOL in SCI/ST and LOI/LE error control modes (SCI/ST, LOI/LE), $p = 5$. SCI/ST and LOI/LE are about twice as fast as BAC/ST and BAC/LE

with an initial condition at $t = 0$ and boundary conditions at $x = 0$ and $x = 1$ taken from the exact solution

$$u(x,t) = \frac{1}{2} - \frac{1}{2}\tanh\left(\frac{x - \frac{t}{2} - \frac{1}{4}}{4\varepsilon}\right),$$

where ε is a problem dependent parameter. In Fig. 1, we compare the BACOL, SCI, and LOI error estimation schemes with the true error, for the OLBE ($\varepsilon = 10^{-3}$) at $t = 1$, with $p = 4$ and a tolerance, $tol = 10^{-4}$. (BACOL error estimates control the mesh.)

We see that the error estimates are generally in good agreement with the true error except for the SCI scheme on subintervals where an adjacent subinterval ratio is large. However, this issue is less significant when the SCI error estimates are used to control the mesh. See additional results in [2].

In Fig. 2, we compare BACOL in ST and LE modes (BAC/ST, BAC/LE) with BACOLI in SCI/ST and LOI/LE modes (SCI/ST, LOI/LE) with respect to efficiency. We again consider the OLBE ($\varepsilon = 10^{-3}$) with final time $t = 1$. We consider $p = 5$ and a set of 91 tol values over the range from 10^{-1} to 10^{-11}.

We see that BAC/ST and BAC/LE have comparable execution times and that these times are significantly greater than the BACOLI code in either SCI/ST or LOI/LE mode. However BACOLI in SCI/LOI mode has 14 failures over the 91 test cases. See [10] for additional results. An examination of the *relative* execution times for this problem (see [10]), averaged over $tol = 10^{-4}, 10^{-6}, 10^{-8}$ and $p = 4, \ldots, 11$, gives $\frac{BAC/LE}{BAC/ST} = 0.93$, $\frac{SCI/ST}{BAC/ST} = 0.57$, $\frac{LOI/LE}{BAC/ST} = 0.62$, and $\frac{LOI/LE}{SCI/ST} = 1.16$.

The BACOLI webpage, where the source code for BACOLI, a number of examples, and a Fortran 95 wrapper for the package are posted, is
http://cs.smu.ca/∼muir/BACOLI-3_Webpage.htm.

References

1. Arsenault, T., Smith, T., Muir, P.H.: Superconvergent interpolants for efficient spatial error estimation in 1D PDE collocation solvers. Can. Appl. Math. Q. **17,** 409–431 (2009)
2. Arsenault, T., Smith, T., Muir, P.H., Keast, P.: Efficient interpolation-based error estimation for 1D time-dependent PDE collocation codes. Saint Mary's University, Dept. of Mathematics and Computing Science Technical Report Series. http://cs.smu.ca/tech_reports/txt2011_001.pdf (2011)
3. Arsenault, T., Smith, T., Muir, P.H., Pew, J.: Asymptotically correct interpolation-based spatial error estimation for 1D PDE solvers. Can. Appl. Math. Q. **20,** 307–328 (2012)
4. Ascher, U.M., Mattheij, R.M.M., Russell, R.D.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations (volume 13 of Classics in Applied Mathematics). Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1995)
5. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations (volume 14 of Classics in Applied Mathematics). Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1996)
6. de Boor, C.: A Practical Guide to Splines, volume 27 of Applied Mathematical Sciences. Springer, New York (1978)
7. Douglas, J. Jr., Dupont, T.: Collocation Methods for Parabolic Equations in a Single Space Variable (Lecture Notes in Mathematics), vol. 385. Springer, Berlin (1974)
8. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II, 2nd edn (volume 14 of Springer Series in Computational Mathematics). Springer, Berlin (1996)
9. Moore, P.K.: Interpolation error-based a posteriori error estimation for two-point boundary value problems and parabolic equations in one space dimension. Numer. Math. **90**(1):149–177 (2001)
10. Pew, J., Li, Z., Muir, P.H.: A computational study of the efficiency of collocation software for 1D parabolic PDEs with interpolation-based spatial error estimation. Saint Mary's University, Dept. of Mathematics and Computing Science Technical Report Series, http://cs.smu.ca/tech_reports/txt2013_001.pdf (2013)
11. Wang, R., Keast, P., Muir, P.H.: BACOL: B-spline Adaptive COLlocation software for 1D parabolic PDEs. ACM Trans. Math. Softw. **30**(4):454–470 (2004)
12. Wang, R., Keast, P., Muir, P.H.: A comparison of adaptive software for 1D parabolic PDEs. J. Comput. Appl. Math. **169**(1):127–150 (2004)
13. Wang, R., Keast, P., Muir, P.H.: Algorithm 874: BACOLR—spatial and temporal error control software for PDEs based on high-order adaptive collocation. ACM Trans. Math. Softw. **34**(3):Art. 15, 28 (2008)

Downscaling of Regional Climate Scenarios within Agricultural Areas Across Canada with a Multivariate, Multisite Model

Nathaniel K. Newlands, Weixun Lu and Tracy A. Porcelli

Abstract Better methods are needed to statistically downscale climate variability to agricultural ecosystem impact scales and to reduce uncertainty in regional climate model (RCM) predictions. We present a multivariate, multisite model for downscaling climate to the 10 km scale for agricultural areas across Canada. Scenario data was obtained from NARCCAP (North American Regional Climate Change Assessment Program). This method employs variable-selection for a multivariate set of regional climate model predictors, and may offer a rapid (automated) and reliable (cross-validated) way to generate high-resolution climate surfaces for use in agricultural decision-making. We provide selected results that show the model can significantly reduce bias in mean precipitation.

1 Introduction

Crop yield forecasting and the integrated assessment of environmental and economic risks of agricultural production both require detailed information on historical and future impacts and variability of climate trends to reliably capture the broad spectrum of potential cumulative impacts of a changing global climate on soil, water and air quality. Typically, higher-resolution downscaled climate information (1–10 km, daily) is required by agroecosystem models and operational monitoring support systems to guide agricultural decision-making. Better methods are needed to statistically downscale climate variability to agricultural ecosystem impact scales and to reduce

N. K. Newlands (✉)
Science and Technology, Agriculture and Agri-Food Canada, Lethbridge Research Centre,
P.O. Box 3000, Lethbridge, AB T1J 4B1, Canada
e-mail: nathaniel.newlands@agr.gc.ca

W. Lu
Department of Geography, University of Victoria, P.O. Box 3060, Station CSC, Victoria,
BC V8W 3R4, Canada
e-mail: lu@uvic.ca

T. A. Porcelli
Physicist/Scientific Consultant, 316 7th Avenue South, Lethbridge,
AB T1J 1H7, Canada

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_48

uncertainty in regional climate model (RCM) predictions [1, 2]. At the 10 km scale, changes in orography, large water bodies, land vegetation cover and other evapotranspiration land-air feedbacks enact a strong regional-scale influence on seasonal changes in climate. Yet, downscaling models often rely on a single predictor variable and generate predictions at single sites without incorporating finer-scale physical influences on climate that change the spatial covariance of precipitation, temperature and other climate variables [3]. In this chapter, we present a multivariate, multisite model for downscaling climate to the 10 km scale for agricultural areas across Canada. Scenario data was obtained from NARCCAP (North American Regional Climate Change Assessment Program) [4]. This method employs variable-selection for a multivariate set of regional climate model predictors. We provide selected results that show the model can significantly reduce bias in mean precipitation.

2 Methodology

2.1 Multisite and Multivariate Selection

A set of representative agricultural areas were selected for validating the statistical downscaling model (Fig. 1), based on the following considerations: (1) span the major agricultural activities and crops, (2) be situated within agricultural land census ecumene boundary, (3) primarily contain agricultural land (Agricultural Land Cover for Canada, circa 2000), (4) reside fully within distinct eco-zones, so that each region is distinguishable based on climate conditions/characteristics (Ecological Framework for Canada), (5) contain a maximum number of high-quality, long-term climate monitoring stations (Meteorological Service of Canada), supplemented with medium quality stations where needed to increase climate time-series data, (6) contain a maximum number of climate model scenario evaluation grid points and (7) contain a maximum number of historical climate interpolation grid points. These data quality and sample size considerations were evaluated using geospatial cross-referencing, layer intersect, other spatial analysis routines provided by ESRI^{TM} ArcGIS^{TM} (Version 10). We superimposed digital elevation satellite imagery (Landsat X, 2010, GoogleEarth^{TM}) on spatially-referenced maps to assess the proportion of agricultural land, number of 50 km regional climate model scenario nodes and climate, elevation and land-form characteristics of the 7 validation areas.

The technique used to obtain estimates from the imagery data was point-based extraction. The size of these areas ranged from 1530 to 46,712 km^2. The number of scenario points ranged between 1 to a maximum of 23, with 14-481 historical climate reference grid-points (10 km grid). Scenario node locations that were situated in each of the geo-referenced validation agricultural regions were determined by point-referencing in ArcGIS. Scenario data at the X, Y locations identified was then extracted from the 50 km resolution gridded data of CGCM3–CRCM3 model (Coupled Canadian Global Climate Model Version 3 with Regional Climate Model Version 3 boundary conditions) output, using available netCDF extraction routines

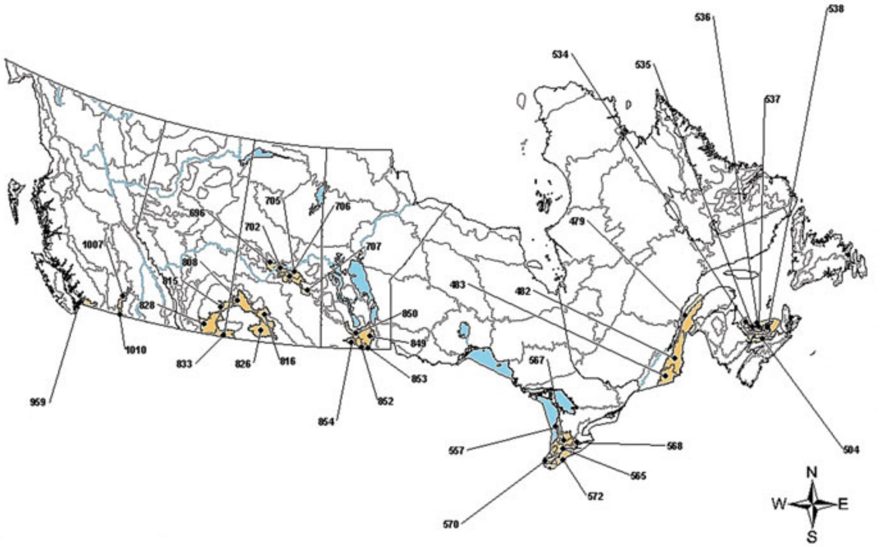


Fig. 1 Validation areas (ecodistricts) representative of different climate, landscape and soil zones

that call a given X, Y coordinate. Climate stations were selected by defining a radial buffer around each scenario location. Within each ecodistrict, an initial number of sites for the nearest-neighbour sampling was determined by selecting a set of historical observation stations with high quality data (i.e. sufficient length of data record and minimal data gaps) that are located nearby every regional climate model scenario point in each validation region (Fig. 2). The station selection and maximum radius of influence was optimized based on separation distance criterion only, using a K-means clustering algorithm. This determined the minimum radius around all the scenario points that lie within a given validation region, whereby each radial zone contained a relatively homogeneous number of stations, without any overlap. The K-nearest neighbours algorithm uses this initial multisite selection (prior distribution) to generate a conditional probability distribution function. This provides enhanced spatio-temporal continuity in the model simulations, by referencing more than just one climate station neighbouring the scenario reference location. Combining more stations than just one provides more input data of measured climate variability across the area of interest. The multisite station data is then used to identify other atmospheric variables that significantly influence the temperature field and rainfall spatial distribution.

The model considers an extended set of atmospheric variables (as higher-order predictors of regional climate variability) and the ranking of their importance or relative influence on precipitation and temperature, such as surface wind speed, surface specific humidity, sea level pressure, incident and reflected shortwave radiation, sensible and latent heat flux, surface evaporation, geopotential height (500 hpa) and atmospheric boundary layer thickness.

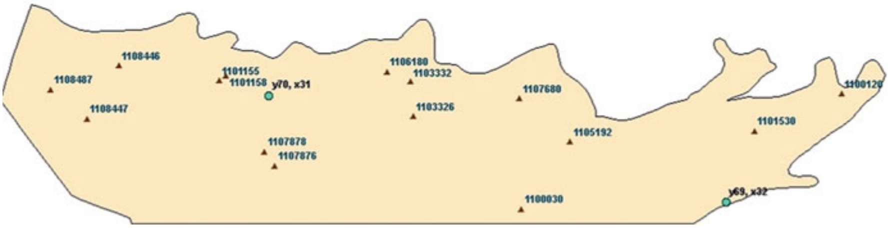


Fig. 2 British Columbia's Fraser Valley (region 1), showing the distribution of observation climate stations (*red* with *blue* IDs), and regional climate model (RCM) scenario downscaling location (*green* with *black* IDs)

2.2 Statistical Model

The conditional distribution $p(R|X)$ can be estimated using kernel density estimation (KDE) or nearest neighbours bootstrap resampling [5]. Here we apply the nearest neighbour approach that can be applied both to linear and non-linear relationships between predictor variables. We define R_t as the climate (e.g. rainfall) response or predictands vector, and A_t a vector of atmospheric predictor variables at time t. X_t is a feature vector of predictor variables (atmospheric or other indices) at time t. The conditional cumulative distribution function (CDF) is given by,

$$p(R_t|X_t) = \frac{p(R_t, X_t)}{p(X_t)} \tag{1}$$

The conditional probability can be expressed as a sum of weighted probabilities associated with an observation at time i, based on a measure of proximity of X_t to X_i (or X_j), given by,

$$p(R_t|X_t) = \sum_i p_i = \sum_i \frac{\psi(X_t - X_i)}{\sum_j \psi(X_j - X_i)} \tag{2}$$

We specify an inverse spatial correlation (proximity) function, with k as the number of observations with Euclidean distance to X_t being less than or equal to the distance between X_t and X_i in the historical data, and K is the maximum value of k,

$$\psi(X_t - X_i) = \begin{cases} \frac{1}{k} & k \leq K \\ 0 & k > K \end{cases} \tag{3}$$

The distance between X_t and X_i is,

$$\delta_{t,i} = \sqrt{\sum_{j=1}^{m} (s_j \beta_j (X_{j,i} - X_{j,t}))^2} \tag{4}$$

where X_i consists of m predictor variables $X_{j,i}$, $j = (1, ..., m)$, and s_j is the scaling weight and β_j, the influence weight associated with the jth predictor. The downscaling method uses least-angle regression (LARS) to optimize the selection of variables to be included in the feature vector (X). The set of model parameters are: the scaling weight vector (s), number of nearest neighbours (K), influence weights (β) and length of the moving window (days) (τ). The estimation of K and τ is data and location specific, and is obtained via sensitivity analysis. The influence weights specify the relative influence each predictor variable has on the CDF, and are determined via leave-one-out cross validation to minimize residual error. Scaling weights were specified as the reciprocal of the sample standard deviation of each predictor variable.

3 Numerical Results

Selected results presented demonstrate that our model corrects significant error in the seasonal-averaged scenario distribution of precipitation in regional climate model output. The change in summer precipitation pattern, hindcasted or backcasted for 1971–2000 having large spatial variability (-1.0 to $+0.5$ mm/day) across different regions of Canada. Summer rainfall in the southern Canadian Prairies and Ontario is predicted to decrease, relative to recent historical levels. Results obtained indicate that rainfall in the Fraser Valley is significantly correlated to previous day rainfall, incident radiation, geo-potential height, sea-level pressure, whereas, in the Atlantic Maritime, sensible surface heat flux, specific humidity and wind speed explain regional rainfall variance.

Downscaled predictions from the model of seasonally-averaged CDFs for precipitation indicate improved accuracy with 6 total predictors within the Fraser Valley region, with $K = 4$ nearest neighbours and $\tau = 3$ days (Fig. 3). The model corrects bias in the CGCM3-CRCM3 summer and winter predictions, especially important when higher amounts of precipitation typically occur.

4 Summary

Results show the statistical downscaling model successfully corrects bias in the statistical distribution moments of daily, monthly, seasonal and annual climate variables (i.e. temperature and precipitation) using addititional climate predictors and historical climate reference data. These findings indicate that 4 neighbours, a window size of 3 days and 6 atmospheric variable predictors can significantly reduce bias in mean precipitation. A sensitivity analysis of higher-order atmospheric variables and subgrid spatial correlation will help to further explain spatial variance in the observed climate trend and improve the accuracy of scenario data at finer (i.e. 10 km) spatial resolutions. This work will provide corrections of regional climate scenario data in generating more reliable forecasts of agricultural crop yield and production across Canada.

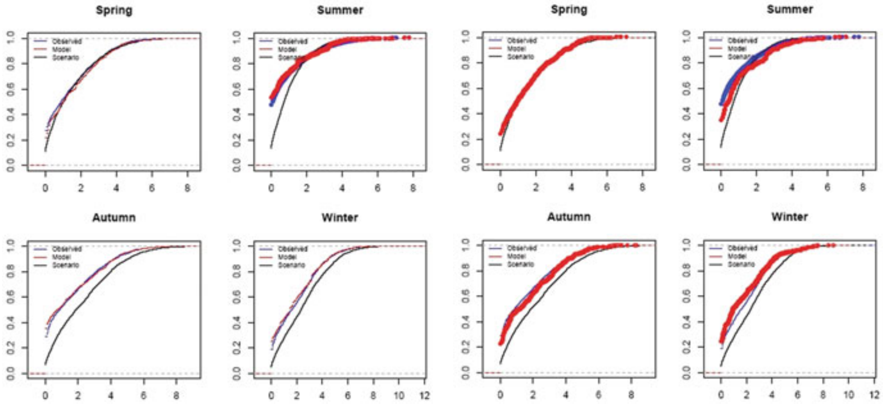


Fig. 3 Downscaling of CGCM3-CRCM3 output for precipitation (x 0.01 mm units) in the Fraser Valley, BC. The *blue line* refers to historical observational data, *black line* is based on 50 km scenario output, and *red line* is the predicted downscaling model. $K = 4$ neighbours, $\tau = 3$ days. Cross-validated mean-squared error (MSE) decreases from 0.137 to 0.027 mm for 1 and 6 predictors, respectively

Acknowledgments We wish to thank (NARCCAP) for providing the data used in this chapter. We also thank the Sustainable AGriculture Environmental Systems (SAGES) Research Program of Agriculture and Agri-Food Canada (AAFC) for funding support. We thank Mr. Abhishek Malwankar (Co-operative Internship Program, University of Lethbridge) for climate scenario data manipulation and processing, and R coding assistance.

References

1. Dibike, Y.B., Gachon, P., St-Hilaire, A., Ouarda, T.B.M.J., Nguyen, Van T.-V.: Uncertainty analysis of statistically downscaled temperature and precipitation regimes in Northern Canada. Theor. Appl. Climatol. **91**, 149–170 (2008)
2. Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themebl, M., Venema, V., Chun, K.P., Goodess, C.M., Jones, R.G., Onof, C., Vracm, M., Thiele-Eich, I.: Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user. Rev. Geophys. **777**, 1–38 (2010)
3. Wilby, R.L., Wigley, T.L.M.: Precipitation predictors for downscaling observed and general circulation model relationships. Int. J. Climatol. **20**, 641–661 (2000)
4. Mearns, L.O., et al.: The North American Regional Climate Change Assessment Program (NARCCAP) dataset, National Centre for Atmospheric Research Earth System Grid data portal, Boulder, CO. (2007, updated 2011, data downloaded 2012–13)
5. Mehrotra, R., Sharma, A.: Conditional resampling of hydrological time series using multiple predictor variables: a K-nearest neighbour approach. Adv. Water Res. **29**, 987–999 (2006)

Iterative Techniques for Nonlinear Periodic Boundary Value Problems (PBVPs) via Initial Value Problems

David H. Dezern and Sudhakar G. Pandit

Abstract We develop constructive methods for solving periodic boundary value problems (PBVPs) associated with a nonlinear first order scalar differential equation in a unified setting. The method of generalized quasilinearization which we employ yields rapid convergence of monotone iterates to the solution of the PBVP. The monotone iterates in our approach are solutions of linear initial value problems (IVPs) as opposed to the linear PBVPs which appear in conventional methods. We provide graphical and numerical illustrations of our results.

1 Introduction

For $J = [0, T]$, $T > 0$, consider the periodic boundary value problem (PBVP)

$$u' = f(t, u) + g(t, u), \quad u(0) = u(T), \quad t \in J. \tag{1}$$

We employ the IVP approach to develop a generalized quasilinear technique for the PBVP (1). This new approach allows us to dispense with some of the stringent conditions required in the conventional approach, wherein the monotone iterates are solutions of appropriate linear PBVPs (see [3]). We shall use natural lower–upper solutions in the development of our techniques. With suitable modifications, coupled lower–upper solutions can also be used to construct the monotone iterates. Finally, we provide numerical and graphical illustrations which indicate that the rate of convergence of the iterates in our methods is more rapid than in the monotone

Dedication The authors dedicate this chapter to the memory of the late Professor V. Lakshmikantham.

D. H. Dezern (✉) · S. G. Pandit
Department of Mathematics, Winston-Salem State University,
Winston-Salem, NC 27110, USA
e-mail: dezernd@wssu.edu

S. G. Pandit
e-mail: pandits@wssu.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_49

iterative technique developed earlier in [4]. See [2] for the details about monotone iterative techniques for a variety of nonlinear problems. Some recent developments about monotone iterative techniques for unified problems are given in [1, 5, 6]. Antiperiodic problems are discussed in [7]. We shall suppress the details in the proof of our main result.

2 Main Result

We begin with the following definition.

Definition 1 Functions $v, w \in C^1[J, \mathbb{R}]$, where $v \le w$ on J, are said to be *natural lower–upper solutions* relative to the PBVP (1) on J if

$$\begin{aligned} v' &\le f(t,v) + g(t,v) \quad v(0) \le v(T), \quad t \in J; \\ w' &\ge f(t,w) + g(t,w) \quad w(0) \ge w(T), \quad t \in J. \end{aligned}$$

For v_0, $w_0 \in C^1[J, \mathbb{R}]$ with $v_0 \le w_0$ on J, we define the *sector* Ω bounded by the functions v_0 and w_0:

$$\Omega = \left\{ u \in C^1[J, \mathbb{R}] : v_0(t) \le u(t) \le w_0(t),\ t \in J \right\}.$$

Our main result below concerns the PBVP (1).

Theorem 1 *Assume that*

(i) v_0, $w_0 \in C^1[J, \mathbb{R}]$, *where* $v_0 \le w_0$ *on* J, *with* v_0 *and* w_0 *being natural lower–upper solutions relative to the PBVP (1) on* J

(ii) f, $g \in C^2[J \times \mathbb{R}, \mathbb{R}]$, *with* $f_{uu} \ge 0$ *and* $g_{uu} \le 0$ *for* $t \in J$ *and* $u \in \Omega$

Then there exist two convergent sequences in Ω, *a nondecreasing sequence* $\{v_n(t)\}$ *and a nonincreasing sequence* $\{w_n(t)\}$ *such that* $v_n \to v$, $w_n \to w$ *uniformly on* J, *where* v *and* w *are the minimal and the maximal solutions, respectively, of the PBVP (1) on* J.

Further, if f *and* g *satisfy either of the following conditions:*

(C$_1$) $f(t,u_1) - f(t,u_2) \le -M_1(u_1 - u_2)$ *and* $g(t,u_1) - g(t,u_2) \ge -M_2(u_1 - u_2)$ *for* $t \in J$ *whenever* $u_1 \ge u_2$ *with constants* M_1 *and* M_2 *such that* $M_1 > M_2 > 0$*; or*

(C$_2$) $\int_0^T [f_u(s, \alpha(s)) - g_u(s, \beta(s))]ds \ne 0$ *for any* α, $\beta \in \Omega$*;*

then $v \equiv w \equiv u$ *and consequently the PBVP (1) has a unique solution in* Ω.

Proof Our proof, which is constructive, differs from the conventional one (see [3]). For $n = 1, 2, 3, \ldots$, define the iterates

$$v_n(t) = e^{-D_{n-1}(t)} \left(v_{n-1}(T) + \int_0^t B_{n-1}(s) \cdot e^{-D_{n-1}(s)} ds \right)$$

and

$$w_n(t) = e^{-D_{n-1}(t)} \left(w_{n-1}(T) + \int_0^t C_{n-1}(s) \cdot e^{-D_{n-1}(s)} ds \right)$$

where

$$A_n(t) = -f_u(t, v_n(t)) - g_u(t, v_n(t));$$

$$B_n(t) = f(t, v_n(t)) + g(t, v_n(t)) + A_n(t)v_n(t);$$

$$C_n(t) = f(t, w_n(t)) + g(t, w_n(t)) + A_n(t)w_n(t);$$

$$\text{and} \quad D_n(t) = \int_0^t A_n(s)ds.$$

By the Ascoli–Arzelà theorem, it follows that there exist limit functions v and w in Ω such that $\lim_{n\to\infty} v_n(t) = v(t)$ and $\lim_{n\to\infty} w_n(t) = w(t)$ uniformly, monotonically, and quadratically, where v and w are the minimal and the maximal solutions, respectively, of the PBVP (1) and $v \leq w$ on J. To establish uniqueness of v and w, it remains to show that $w \leq v$ on J. To this end, let $p(t) = w(t) - v(t)$, and suppose that condition (C$_1$) holds. Then $p'(t) = w'(t) - v'(t) \leq -Mp(t)$, where $M = M_1 - M_2$. This yields $p(t) \leq p(0) \cdot e^{-Mt}$ for $t \in J$. Setting $t = T$ we obtain $p(0) = p(T) \leq p(0) \cdot e^{-MT}$. Since M and T are both positive, this implies that $p(0) \leq 0$ and consequently $p(t) \leq 0$ on J.

If condition (C$_2$) holds, then by the mean value theorem we have $p'(t) = f_u(t, \alpha(t)) \cdot p(t)$ for $t \in J$ for some $\alpha \in \Omega$ (for which $v_0 \leq \alpha \leq w_0$). This implies $p(t) = p(0) \cdot \exp\left(\int_0^t f_u(s, \alpha(s))ds\right)$. Setting $t = T$ we obtain, as before, that $p(0) = p(T) \cdot \exp\left(\int_0^T f_u(s, \alpha(s))ds\right)$. By condition (C$_2$), this implies $p(0) = 0$ and hence $p(t) \equiv 0$ on J, which completes the proof of the theorem. □

3 An Illustrative Example

Here we provide an example to illustrate Theorem 1. For $J = [0, 1]$ consider the unified PBVP

$$u' = f(t, u) + g(t, u), \quad u(0) = u(1), \quad t \in J, \tag{2}$$

where

$$f(t, u) = \begin{cases} -1.25 \sin u & \text{if } 0 \leq u \leq \frac{\pi}{2} \\ -1.25 & \text{if } u > \frac{\pi}{2} \\ 0 & \text{if } u < 0 \end{cases}$$

and

$$g(t, u) = \begin{cases} \cos u & \text{if } 0 \leq u \leq \frac{\pi}{2} \\ 0 & \text{if } u > \frac{\pi}{2} \\ 1 & \text{if } u < 0. \end{cases}$$

Table 1 Monotone iterative technique

n	$u(t)-v_n(t)$	$w_n(t)-u(t)$	$w_n(t)-v_n(t)$
1	0.312510	0.443268	0.755778
2	0.152541	0.221977	0.374518
3	0.070070	0.103175	0.173245
4	0.031330	0.046227	0.077557
5	0.014027	0.020697	0.034724
6	0.006292	0.009285	0.015577
7	0.002821	0.004164	0.006986
8	0.001265	0.001867	0.003132
9	0.000567	0.000838	0.001404
10	0.000254	0.000376	0.000630

Table 2 Generalized quasilinear technique

n	$u(t)-v_n(t)$	$w_n(t)-u(t)$	$w_n(t)-v_n(t)$
1	0.312510	0.443268	0.755778
2	0.144803	0.211611	0.356414
3	0.053225	0.079179	0.132404
4	0.015224	0.022731	0.037955
5	0.003618	0.005407	0.009025
6	0.000771	0.001152	0.001922
7	0.000157	0.000236	0.000393
8	0.000032	0.000048	0.000080
9	0.000006	0.000010	0.000016
10	0.000001	0.000002	0.000003

Let $v_0(t) \equiv 0$ and $w_0(t) \equiv \frac{\pi}{2}$. Then v_0 and w_0 form a pair of natural lower–upper solutions for the PBVP (2) and we have $f_{uu} \geq 0$ and $g_{uu} \leq 0$ on Ω. For $n = 1, 2, 3, \ldots$, define v_n and w_n as in Theorem 1. For $t = 0.75$, numerical results are shown in Table 2 above. For the same value of t, Table 1 shows corresponding results given by the monotone iterative technique, in which the mode of convergence of the iterates is linear. Graphical results for the generalized quasilinear technique and the monotone iterative technique are displayed in Fig. 1 and Fig. 2, respectively. It is readily seen that the convergence of iterates in the generalized quasilinear technique is more rapid than in the monotone iterative technique.

We note in passing that the exact solution of the PBVP (2) is $\tan^{-1}\left(\frac{4}{5}\right)$.

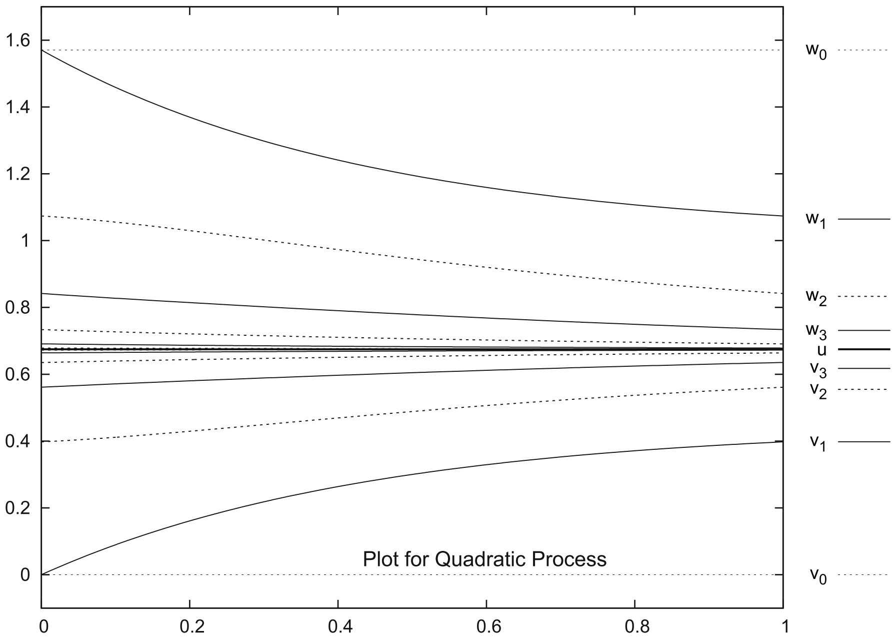


Fig. 1 The figure includes plots of the natural lower–upper solutions v_0, w_0 for the PBVP (2) together with the first ten iterates v_n, w_n generated by the generalized quasilinear technique, as well as a plot of the known exact solution $u = \tan^{-1}(\frac{4}{5})$. The plots of the iterates are alternately solid or dashed, and to reduce clutter, explicit references to iterates v_n or w_n for $n > 3$ have been removed from the legend, even though the points on the graphs have faithfully been plotted

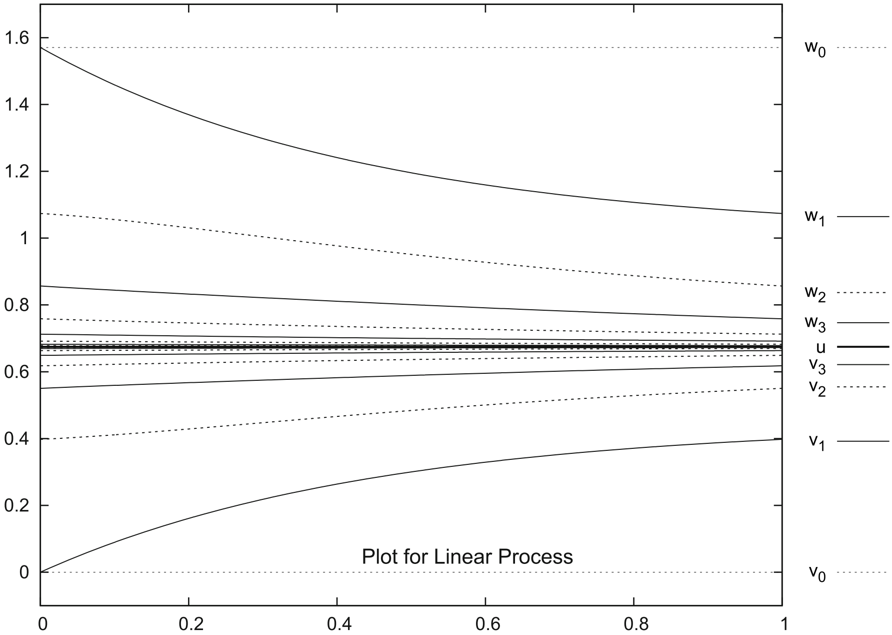


Fig. 2 The figure includes plots of the natural lower–upper solutions v_0, w_0 for the PBVP (2) together with the first ten iterates v_n, w_n generated by the monotone iterative technique, as well as a plot of the known exact solution $u = \tan^{-1}(\frac{4}{5})$. The plots of the iterates are alternately *solid* or *dashed*, and to reduce clutter, explicit references to iterates v_n or w_n for $n > 3$ have been removed from the legend, even though the points on the graphs have faithfully been plotted

References

1. Bhaskar, T.G., McRae, F.A.: Monotone iterative techniques for nonlinear problems involving the difference of two monotone functions. Appl. Math. Comput. **133**(1):187–192 (2002)
2. Ladde, G.S., Lakshmikantham, V., Vatsala, A.S.: Monotone Iterative Techniques for Nonlinear Differential Equations. Pitman, Boston (1985)
3. Lakshmikantham, V., Vatsala, A.S.: Generalized Quasilinearization for Nonlinear Problems. Kluwer, Dordrecht (1998)
4. Pandit, S.G., Dezern, D.H., Adeyeye, J.O.: A new approach to monotone iterative techniques for nonlinear periodic boundary value problems. Proc. Dyn. Syst. Appl. **6,** 303–309 (2012)
5. Sokol, M., Vatsala, A.S.: A unified exhaustive study of monotone iterative method for initial value problems. Nonlinear Stud. **8**(4):429–438 (2001)
6. West, I.H., Vatsala, A.S.: Generalized monotone iterative method for initial value problems. Appl. Math. Lett. **17**(11):1231–1237 (2004)
7. Yin, Y.: Monotone iterative technique and quasilinearization for some anti-periodic problems. Nonlinear World **3**(2):253–266 (1996)

Fast and Stable Algorithms for Discrete Sine Transformations having Orthogonal Factors

Sirani M. Perera and Vadim Olshevsky

Abstract In this chapter we derive fast, recursive, and numerically stable radix-2 algorithms for discrete sine transformations (DST) having sparse and orthogonal factors. These real radix-2 stable algorithms are completely recursive, fast, and based on the simple orthogonal factors. Comparing to the known bulky and mostly unstable DST algorithms, our algorithms are easy to implement and use only permutations, scaling by constants, butterfly operations, and plane rotations/rotation-reflections.

For a given vector $\mathbf{x}$, we also analyze error bounds of computing $\mathbf{y} = S\mathbf{x}$ for the presented DST algorithms: S. Finally a classification of these real radix-2 DST algorithms enables us to establish the excellent forward and backward stability based on the sparse and orthogonal factors.

1 Introduction

Discrete Fourier transforms (DFT) have numerous applications in sciences and engineering especially in applied mathematics and electrical engineering. There are real versions of the DFT called the discrete sine transform and the discrete cosine transform of main variants I–IV, and they have been studied by several authors, see, e.g., [1–5].

The four main variants of DST can be denoted by

$$S^{I}_{n-1} = \sqrt{\tfrac{2}{n}}\left[\sin\tfrac{(j+1)(k+1)\pi}{n}\right]_{j,k=0}^{n-2}, \quad S^{II}_{n} = \sqrt{\tfrac{2}{n}}\left[\varepsilon_n(j+1)\sin\tfrac{(j+1)(2k+1)\pi}{2n}\right]_{j,k=0}^{n-1},$$

$$S^{III}_{n} = \left[S^{II}_{n}\right]^{T}, \qquad S^{IV}_{n} = \sqrt{\tfrac{2}{n}}\left[\sin\tfrac{(2j+1)(2k+1)\pi}{4n}\right]_{j,k=0}^{n-1}$$

S. M. Perera (✉)
Daytona State College, Daytona Beach, FL 32114, USA
e-mail: pereras@daytonastate.edu

V. Olshevsky
University of Connecticut, Storrs, CT 06269, USA
e-mail: olshevsky@math.uconn.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_50

where $\varepsilon_n(0) = \varepsilon_n(n) = \frac{\sqrt{2}}{2}$, $\varepsilon_n(j) = 1$ for $j \in \{1, 2, \cdots, n-1\}$ and $n \geq 2$ is an integer. In Sect. 2 we use a permutation matrix and trigonometric identities to derive orthogonal and sparse factors for DST I–IV. We state algorithms for DST I–IV and declare that the cost is $O(nlogn)$ operations in Sect. 3. In Sect. 4 we show that our factors for DST I–IV are numerically stable and derive error bounds.

2 Sparse and Orthogonal Factors for DST I–IV

This section introduces a complete factorization for DST I–IV having sparse, orthogonal, rotation/rotation-reflection, and butterfly matrices. Before deriving the factorizations let's introduce a vector in $\mathbb{R}^n$ by $\mathbf{x} = \left[x_0, x_1, \cdots, x_{n-1}\right]$, an involution matrix $\tilde{I}_n$ by $\tilde{I}_n \mathbf{x} = \left[x_{n-1}, x_{n-2}, \cdots, x_0\right]^T$, a block diagonal matrix, i.e., blkdiag(M, N) by $[M \odot N]$ and, for $n \geq 3$ an even–odd permutation matrix P_n by

$$P_n \mathbf{x} = \begin{cases} \left[x_0, x_2, \cdots, x_{n-2}, x_1, x_3, \cdots, x_{n-1}\right]^T & \text{for even}\, n, \\ \left[x_0, x_2, \cdots, x_{n-1}, x_1, x_3, \cdots, x_{n-2}\right]^T & \text{for odd}\, n. \end{cases}$$

Lemma 1 *Let $n \geq 4$ be an even integer. The matrix S_n^{II} can be factored in the form*

$$S_n^{II} = P_n^T \left[S_{\frac{n}{2}}^{IV} \odot S_{\frac{n}{2}}^{II}\right] H_n \, where \, H_n = \frac{1}{\sqrt{2}} \begin{bmatrix} I_{\frac{n}{2}} & \tilde{I}_{\frac{n}{2}} \\ I_{\frac{n}{2}} & -\tilde{I}_{\frac{n}{2}} \end{bmatrix}. \tag{1}$$

Proof Permuting the rows in S_n^{II} gives

$$\sqrt{\frac{2}{n}} \left[\begin{array}{c|c} \left[\sin \frac{(2j+1)(2k+1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} & \left[\sin \frac{(2j+1)(n+2k+1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} \\ \hline \left[\varepsilon_n(2j+2) \sin \frac{(2j+2)(2k+1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} & \left[\varepsilon_n(2j+2) \sin \frac{(2j+2)(n+2k+1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} \end{array} \right]$$

$$= \sqrt{\frac{2}{n}} \left[\begin{array}{c|c} \left[\sin \frac{(2j+1)(2k+1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} & \left[\sin \frac{(2j+1)(n-2k-1)\pi}{2n}\right]_{j,k=0}^{\frac{n}{2}-1} \\ \hline \left[\varepsilon_{\frac{n}{2}}(j+1) \sin \frac{(j+1)(2k+1)\pi}{n}\right]_{j,k=0}^{\frac{n}{2}-1} & -\left[\varepsilon_{\frac{n}{2}}(j+1) \sin \frac{(j+1)(n-2k-1)\pi}{n}\right]_{j,k=0}^{\frac{n}{2}-1} \end{array} \right]$$

$$= \frac{1}{\sqrt{2}} \left[\begin{array}{c|c} S_{\frac{n}{2}}^{IV} & S_{\frac{n}{2}}^{IV} \tilde{I}_{\frac{n}{2}} \\ \hline S_{\frac{n}{2}}^{II} & -S_{\frac{n}{2}}^{II} \tilde{I}_{\frac{n}{2}} \end{array} \right] = \left[S_{\frac{n}{2}}^{IV} \odot S_{\frac{n}{2}}^{II}\right] H_n.$$

Lemma 2 *Let $n \geq 4$ be an even integer. The matrix S^I_{n-1} can be factored in the form*

$$S^I_{n-1} = P^T_{n-1}\left[S^{III}_{\frac{n}{2}} \odot S^I_{\frac{n}{2}-1}\right]\widehat{H}_{n-1} \; where \; \widehat{H}_{n-1} = \frac{1}{\sqrt{2}}\begin{bmatrix} I_{\frac{n}{2}-1} & & \widetilde{I}_{\frac{n}{2}-1} \\ & \sqrt{2} & \\ I_{\frac{n}{2}-1} & & -\widetilde{I}_{\frac{n}{2}-1} \end{bmatrix}.$$

Proof The proof of lemma 2 follows the silmilar lines as the proof of lemma 1.

Remark 1 By the transpose of (1), DST III is given by $S^{III}_n = H^T_n\left[\, S^{IV}_{\frac{n}{2}} \odot S^{III}_{\frac{n}{2}}\,\right]P_n$.

Proposition 1 *Let $n \geq 2$ be an integer. Then:*

$$\tilde{I}_n C^{II}_n = S^{II}_n D_n, D_n S^{II}_n = S^{II}_n \tilde{I}_n, D_n C^{II}_n = C^{II}_n \tilde{I}_n \tag{2}$$

where $C^{II}_n = \sqrt{\frac{2}{n}}\left[\varepsilon_n(j) cos\frac{j(2k+1)\pi}{2n}\right]^{n-1}_{j,k=0}$ *and* $D_n = diag\left((-1)^k\right)^{n-1}_{k=0}$.

The straightforward proof of (2) is given by matrix multiplication. One can use (2) and simple trigonometric identities to derive factors for the DST IV.

Lemma 3 *Let $n \geq 4$ be an even integer. The matrix S^{IV}_n can be factored in the form*

$$S^{IV}_n = P^T_n V_n\left[S^{II}_{\frac{n}{2}} \;\odot\; S^{II}_{\frac{n}{2}}\right]Q_n \tag{3}$$

where

$$\begin{aligned} V_n &= \left[1 \odot \frac{1}{\sqrt{2}}\begin{bmatrix} I_{\frac{n}{2}-1} & -I_{\frac{n}{2}-1} \\ -I_{\frac{n}{2}-1} & -I_{\frac{n}{2}-1} \end{bmatrix} \odot -1\right]\left[\widetilde{I}_{\frac{n}{2}} \odot D_{\frac{n}{2}}\right], \\ C_{\frac{n}{2}} &= \left(cos\frac{(2k+1)\pi}{4n}\right)^{\frac{n}{2}-1}_{k=0}, \; S_{\frac{n}{2}} = \left(sin\frac{(2k+1)\pi}{4n}\right)^{\frac{n}{2}-1}_{k=0}, \\ Q_n &= \left[D_{\frac{n}{2}} \odot I_{\frac{n}{2}}\right]\begin{bmatrix} diag S_{\frac{n}{2}} & \left(diag C_{\frac{n}{2}}\right)\widetilde{I}_{\frac{n}{2}} \\ -\widetilde{I}_{\frac{n}{2}}\left(diag C_{\frac{n}{2}}\right) & diag\left(\widetilde{I}_{\frac{n}{2}} S_{\frac{n}{2}}\right) \end{bmatrix}. \end{aligned} \tag{4}$$

Proof Apply the matrix P_n to S^{IV}_n to permute rows

$$P_n S^{IV}_n = \sqrt{\frac{2}{n}}\left[\begin{array}{c|c} \left[\sin\frac{(4j+1)(2k+1)\pi}{4n}\right]^{\frac{n}{2}-1}_{j,k=0} & \left[\sin\frac{(4j+1)(n+2k+1)\pi}{4n}\right]^{\frac{n}{2}-1}_{j,k=0} \\ \hline \left[\sin\frac{(4j+3)(2k+1)\pi}{4n}\right]^{\frac{n}{2}-1}_{j,k=0} & \left[\sin\frac{(4j+3)(n+2k+1)\pi}{4n}\right]^{\frac{n}{2}-1}_{j,k=0} \end{array}\right] \tag{5}$$

By the (1,1) block in (5) and (2)

$$\sqrt{\frac{2}{n}}\left[\sin\frac{j(2k+1)\pi}{n}\cos\frac{(2k+1)\pi}{4n}+\cos\frac{j(2k+1)\pi}{n}\sin\frac{(2k+1)\pi}{4n}\right]_{j,k=0}^{\frac{n}{2}-1} \tag{6}$$
$$=\frac{1}{\sqrt{2}}\left(Z_{\frac{n}{2}}S_{\frac{n}{2}}^{II}(\mathrm{diag}C_{\frac{n}{2}})+I'_{\frac{n}{2}}C_{\frac{n}{2}}^{II}(\mathrm{diag}S_{\frac{n}{2}})\right)$$
$$=\frac{1}{\sqrt{2}}\left(Z_{\frac{n}{2}}D_{\frac{n}{2}}S_{\frac{n}{2}}^{II}\tilde{I}_{\frac{n}{2}}(\mathrm{diag}C_{\frac{n}{2}})+I'_{\frac{n}{2}}C_{\frac{n}{2}}^{II}(\mathrm{diag}S_{\frac{n}{2}})\right)$$

where $I'_{\frac{n}{2}}=\left[\sqrt{2}\quad\odot\quad I_{\frac{n}{2}-1}\right]$, $Z_{\frac{n}{2}}\mathbf{t}=\left[0,t_0,t_1,\cdots,t_{\frac{n}{2}-2}\right]^T$ and $\mathbf{t}=\left[t_j\right]_{j=0}^{\frac{n}{2}-1}$.

Using the same procedure as in (6) with the difference of angles formula for sine and (2), the (2,1) block in (5) can be expressed by

$$\sqrt{\frac{2}{n}}\left[\sin\frac{(4j+3)(2k+1)\pi}{4n}\right]_{j,k=0}^{\frac{n}{2}-1}=\frac{1}{\sqrt{2}}\left(I''_{\frac{n}{2}}D_{\frac{n}{2}}S_{\frac{n}{2}}^{II}\tilde{I}_{\frac{n}{2}}(\mathrm{diag}C_{\frac{n}{2}})-Z_{\frac{n}{2}}^{T}C_{\frac{n}{2}}^{II}(\mathrm{diag}S_{\frac{n}{2}})\right) \tag{7}$$

where $I''_{\frac{n}{2}}=\left[I_{\frac{n}{2}-1}\quad\odot\quad\sqrt{2}\right]$.
The (1,2) block in (5) can be restated as

$$\sqrt{\frac{2}{n}}\left[(-1)^j\sin\frac{j(2k+1)\pi}{n}\sin\frac{(n-2k-1)\pi}{4n}\right.$$
$$\left.+(-1)^j\cos\frac{j(2k+1)\pi}{n}\cos\frac{(n-2k-1)\pi}{4n}\right]_{j,k=0}^{\frac{n}{2}-1} \tag{8}$$
$$=\frac{1}{\sqrt{2}}\left(D_{\frac{n}{2}}Z_{\frac{n}{2}}S_{\frac{n}{2}}^{II}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}S_{\frac{n}{2}}))+D_{\frac{n}{2}}I'_{\frac{n}{2}}C_{\frac{n}{2}}^{II}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}C_{\frac{n}{2}}))\right)$$
$$=\frac{1}{\sqrt{2}}\left(-Z_{\frac{n}{2}}D_{\frac{n}{2}}S_{\frac{n}{2}}^{II}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}S_{\frac{n}{2}}))+I'_{\frac{n}{2}}C_{\frac{n}{2}}^{II}\tilde{I}_{\frac{n}{2}}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}C_{\frac{n}{2}}))\right)$$

Here we used $D_{\frac{n}{2}}I'_{\frac{n}{2}}=I'_{\frac{n}{2}}D_{\frac{n}{2}}$, $-D_{\frac{n}{2}}Z_{\frac{n}{2}}=Z_{\frac{n}{2}}D_{\frac{n}{2}}$ and (2).
Finally, using the same procedure as in (8) with the difference of angles formula for sine, $D_{\frac{n}{2}}I''_{\frac{n}{2}}=I''_{\frac{n}{2}}D_{\frac{n}{2}}$, $-D_{\frac{n}{2}}Z_{\frac{n}{2}}=Z_{\frac{n}{2}}D_{\frac{n}{2}}$ and (2), results in

$$\sqrt{\frac{2}{n}}\left[\sin\frac{(4j+3)(n+2k+1)\pi}{4n}\right]_{j,k=0}^{\frac{n}{2}-1}=-\frac{1}{\sqrt{2}}\left(I''_{\frac{n}{2}}D_{\frac{n}{2}}S_{\frac{n}{2}}^{II}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}S_{\frac{n}{2}}))\right.$$
$$\left.+Z_{\frac{n}{2}}^{T}C_{\frac{n}{2}}^{II}\tilde{I}_{\frac{n}{2}}(\mathrm{diag}(\tilde{I}_{\frac{n}{2}}C_{\frac{n}{2}}))\right) \tag{9}$$

By using (6), (7), (8), (9), and (2)

$$P_n S_n^{IV} = \frac{1}{\sqrt{2}} \begin{bmatrix} I'_{\frac{n}{2}} \tilde{I}_{\frac{n}{2}} & -Z_{\frac{n}{2}} D_{\frac{n}{2}} \\ -Z^T_{\frac{n}{2}} \tilde{I}_{\frac{n}{2}} & -I''_{\frac{n}{2}} D_{\frac{n}{2}} \end{bmatrix} \begin{bmatrix} S^{II}_{\frac{n}{2}} \odot S^{II}_{\frac{n}{2}} \end{bmatrix} \begin{bmatrix} D_{\frac{n}{2}} \left(\mathrm{diag} S_{\frac{n}{2}}\right) & D_{\frac{n}{2}} \widetilde{I}_{\frac{n}{2}} \left(\mathrm{diag}\left(\widetilde{I}_{\frac{n}{2}} C_{\frac{n}{2}}\right)\right) \\ -\widetilde{I}_{\frac{n}{2}} \left(\mathrm{diag} C_{\frac{n}{2}}\right) & \mathrm{diag}\left(\widetilde{I}_{\frac{n}{2}} S_{\frac{n}{2}}\right) \end{bmatrix} \quad (10)$$

where the first matrix with the scaling factor $\frac{1}{\sqrt{2}}$ in the RHS of (10) is the same as V_n in (4). Using $\widetilde{I}_{\frac{n}{2}} \left(\mathrm{diag}\left(\widetilde{I}_{\frac{n}{2}} C_{\frac{n}{2}}\right)\right) = \left(\mathrm{diag} C_{\frac{n}{2}}\right) \widetilde{I}_{\frac{n}{2}}$ results in S_n^{IV}.

Remark 2 Following (4), one can perceive that $V_n V_n^T = Q_n Q_n^T = I_n$. Moreover V_n is a sparse matrix having only two upside V-structures and Q_n is a combination of a rotation and rotation–reflection matrices having a butterfly structure.

3 Fast and Recursive Algorithms for DST I–IV

The complete recursive algorithms for DST II and IV can be obtained by (1) and (3). We state this via **algorithms [1]** and **[2]**. A recursive algorithm for DST III is given by remark 1, **[1]**, and **[2]**. For DST I the recursive algorithm is given by lemma 2, remark 1, **[1]**, and **[2]**.

Let $n_k = 2^{t-k}(t \geq 2)$ where $k(0 \leq k \leq t-1)$ is the step number. By (Fig. 1) when $k = i(1 \leq i \leq t-2)$, each $S^{II}_{n_i}$ and $S^{IV}_{n_i}$ has to be subdivided into $S^{IV}_{n_{i+1}}$, $S^{II}_{n_{i+1}}$ and $S^{II}_{n_{i+1}}$, $S^{II}_{n_{i+1}}$ respectively to produce the total of 2^{i+1} nodes. Continuing (Fig. 1) recursively, one can obtain $O(n\log n)$ algorithm for DST II and, similarly for DST I, III, and IV.

4 Error Bounds and Numerical Stability of DST I–IV

The stability and error bound of computing the matrix-vector product $\mathbf{y} = S\mathbf{x}$ where S stands for S^I_{n-1}, S^{II}_n, S^{III}_n, and S^{IV}_n is the main concern in this section.

Theorem 1 *Let $n = 2^t(t \geq 2)$. The error bound for $\boldsymbol{y} = S^{II}_n \boldsymbol{x}$ is given by*

$$\frac{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_2}{\|\boldsymbol{y}\|_2} \leq \left[(1-5u)^{1-t} - 1\right] \quad (11)$$

where $\widehat{\boldsymbol{y}} = fl(S^{II}_n \boldsymbol{x})$ and u is the unit roundoff.

Proof By recursively applying DST II (1) we obtain

$$S^{II}_n = \left(\mathbf{P}^T_0 \mathbf{V}_0\right) \left(\mathbf{P}^T_1 \mathbf{V}_1\right) \cdots \left(\mathbf{P}^T_{t-2} \mathbf{V}_{t-2}\right) \mathbf{S}_{t-1} \left(\mathbf{H}_{t-2}\right) \left(\mathbf{H}_{t-3}\right) \cdots \left(\mathbf{H}_0\right)$$

Algorithm 1 Discrete Sine Transformation II $(DSTII(\mathbf{x},n))$

```
Input: $n = 2^t (t \geq 1), n_1 = \frac{n}{2}, \mathbf{x} \in \mathbb{R}^n$.
1. If $n = 2$, then
       $\mathbf{y} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}$.
2. If $n \geq 4$, then
       $[u_j]_{j=0}^{n-1} := H_n \mathbf{x}$, $\mathbf{z1} := DSTIV\left([u_j]_{j=0}^{n_1-1}, n_1\right)$, $\mathbf{z2} := DSTII\left([u_j]_{j=n_1}^{n-1}, n_1\right)$,
       $\mathbf{y} := P_n^T \left(\mathbf{z1}^T, \mathbf{z2}^T\right)^T$.
Output: $\mathbf{y} = S_n^{II} \mathbf{x}$.
```

Algorithm 2 Discrete Sine Transformation IV $(DSTIV(\mathbf{x},n))$

```
Input: $n = 2^t (t \geq 1), n_1 = \frac{n}{2}, \mathbf{x} \in \mathbb{R}^n$.
1. If $n = 2$, then
       $\mathbf{y} := \begin{bmatrix} \sin\frac{\pi}{8} & \cos\frac{\pi}{8} \\ \cos\frac{\pi}{8} & -\sin\frac{\pi}{8} \end{bmatrix} \mathbf{x}$.
2. If $n \geq 4$, then
       $[u_j]_{j=0}^{n-1} := Q_n \mathbf{x}$, $\mathbf{z1} := DSTII\left([u_j]_{j=0}^{n_1-1}, n_1\right)$, $\mathbf{z2} := DSTII\left([u_j]_{j=n_1}^{n-1}, n_1\right)$,
       $\mathbf{w} := V_n \left(\mathbf{z1}^T, \mathbf{z2}^T\right)^T$, $\mathbf{y} := P_n^T \mathbf{w}$.
Output: $\mathbf{y} = S_n^{IV} \mathbf{x}$.
```

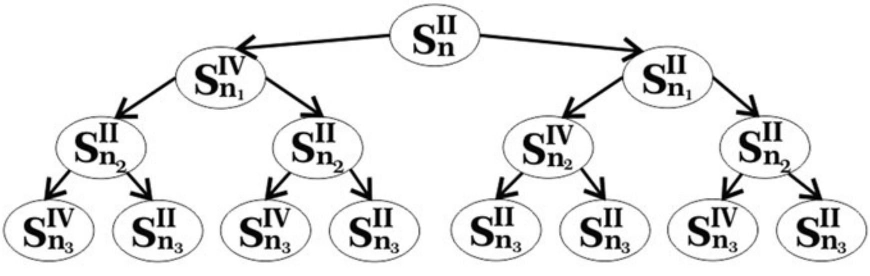


Fig. 1 This shows the first couple of factorization steps of S_n^{II}. The complete factorization for S_n^{II} can be obtained by applying the divide and conquer technique and the cost is only $O(n\log n)$

where

$$\begin{aligned}
&\mathbf{P}_0 := P_n, \mathbf{P}_k := \odot_{2^k}\left\{P_{\frac{n}{2^k}}\right\}, \mathbf{V}_0 := I_n, \mathbf{V}_k := \odot_{2^k}\left\{I_{\frac{n}{2^k}}, V_{\frac{n}{2^k}}\right\},\\
&\mathbf{S}_{t-1} := \odot_{2^{t-1}}\left\{S_2^{II}, S_2^{IV}\right\},\\
&\mathbf{H}_0 := H_n, \mathbf{H}_k := \odot_{2^k}\left\{H_{\frac{n}{2^k}}, Q_{\frac{n}{2^k}}\right\}, \odot_*\{\ \} := \text{no. of block diagonals in the set},\\
&k = 1, 2, \cdots, t-2.
\end{aligned}$$

Using floating point arithmetic for $\mathbf{y} = S_n^{II}\mathbf{x}$

$$\begin{aligned}
\widehat{\mathbf{y}} = &\left(\mathbf{P}_0^T(\mathbf{V}_0 + \Delta\mathbf{V}_0)\right)\left(\mathbf{P}_1^T(\mathbf{V}_1 + \Delta\mathbf{V}_1)\right)\cdots\left(\mathbf{P}_{t-2}^T(\mathbf{V}_{t-2} + \Delta\mathbf{V}_{t-2})\right)(\mathbf{S}_{t-1} + \Delta\mathbf{S}_{t-1})\\
&(\mathbf{H}_{t-2} + \Delta\mathbf{H}_{t-2})(\mathbf{H}_{t-3} + \Delta\mathbf{H}_{t-3})\cdots(\mathbf{H}_0 + \Delta\mathbf{H}_0)\mathbf{x}.
\end{aligned}$$

Each row in $\mathbf{V}_k$ has at most two non zero entries with most 1's and each row in $\mathbf{H}_k$ and $\mathbf{S}_{t-1}$ has at most two non zero entries so

$$\begin{aligned}
&|\Delta\mathbf{V}_0| = 0, |\Delta\mathbf{V}_k| \le \gamma_2\,|\mathbf{V}_k|, |\Delta\mathbf{H}_0| \le \gamma_2\,|\mathbf{H}_0|, |\Delta\mathbf{H}_k| \le \gamma_3\,|\mathbf{H}_k|,\\
&|\Delta\mathbf{S}_{t-1}| \le \gamma_3\,|\mathbf{S}_{t-1}|, \text{for } k = 1, 2, \cdots, t-2.
\end{aligned}$$

Hence

$$\begin{aligned}
|\mathbf{y} - \widehat{\mathbf{y}}| \le &\left[(1+\gamma_5)^{t-1} - 1\right]\mathbf{P}_0^T\,|\mathbf{V}_0|\,\mathbf{P}_1^T\,|\mathbf{V}_1|\cdots\\
&\mathbf{P}_{t-2}^T\,|\mathbf{V}_{t-2}|\,|\mathbf{S}_{t-1}|\,|\mathbf{H}_{t-2}|\,|\mathbf{H}_{t-3}|\cdots|\mathbf{H}_0|\,|\mathbf{x}|.
\end{aligned}$$

By $\gamma_5 := \frac{5u}{1-5u}$, $\|\mathbf{V}_k\|_2 = \|\mathbf{S}_{t-1}\|_2 = \|\mathbf{H}_k\|_2 = 1$ and orthogonality of S_n^{II}(i.e., $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$), we get (11).

Corollary 1 *The error bound for* $\boldsymbol{y} = S_n^{III}\boldsymbol{x}$ *is the same as in (11).*

Theorem 2 *The error bound for* $\boldsymbol{y} = S\boldsymbol{x}$ *is given by* $\frac{\|\mathbf{y}-\widehat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \le \left[(1-5u)^{-t} - 1\right]$ *where* S *stands for* S_{n-1}^I *and* S_n^{IV}, $n = 2^t (t \ge 2)$, *u is the unit roundoff and* $\widehat{\boldsymbol{y}} = fl(S\boldsymbol{x})$.

Proof The proof of theorem 2 follows similar lines as in the proof of theorem 1.

Corollary 2 $\boldsymbol{y} = S\boldsymbol{x}$ *where* S *stands for* S_{n-1}^I, S_n^{II}, S_n^{III}, *and* S_n^{IV} *are forward and backward stable.*

Proof By theorem 1, the radix-2 DST II yields a tiny forward error provided that $\sin\frac{r\pi}{4n}$ and $\cos\frac{r\pi}{4n}$ $(r = 1, 3\cdots, n-1)$ are computed stably. It follows that the computation is backward stable as $\widehat{\mathbf{y}} = S_n^{II}\mathbf{x} + \Delta\mathbf{y} = S_n^{II}(\mathbf{x} + \Delta\mathbf{x})$ with $\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\Delta\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$. If we form $\mathbf{y} = S_n^{II}\mathbf{x}$ by usual multiplication, then $\frac{\|\mathbf{y}-\widehat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \le \gamma_n := \frac{nu}{1-nu}$. Hence S_n^{II} has an error bound smaller than that for usual multiplication by the same factor as the reduction in complexity method, so DST II is perfectly stable. By theorem 2, the stability of DST I and IV follow the same line. Stability of DST III easily follows by the transpose.

References

1. Kailath, T., Olshevsky, V.: Displacement structure approach to discrete trigonometric transform based preconditioners of G. Strang and T. Chan types. SIAM J. Matrix Anal. Appl. **26**, 706–734 (2005)
2. Plonka, G., Tasche, M.: Fast and numerically stable algorithms for discrete cosine transforms. Linear Algebra Appl. **394**, 309–345 (2005)
3. Rao, K.R., Yip, P.: Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic, San Diego (1990)
4. Strang, G.: Introduction to Applied Mathematics. Wellesley-Cambridge Press, Wellesley (1986)
5. Van Loan, C.: Computational Frameworks for the Fast Fourier Transform. SIAM Publications, Philadelphia (1992)

Interactive Computational Search Strategy of Periodic Solutions in Essentially Nonlinear Dynamics

Lev F. Petrov

Abstract We consider essentially nonlinear autonomous and nonautonomous dynamic systems described by ordinary differential equations. In such systems, for the same parameters of the system and forcing, different stable and unstable periodic solutions of different periods can exist. In addition, along with the ordered movements, the existence of a strange attractor is known. In such circumstances, the search for periodic solutions and their stability analysis is not a trivial problem. In order to find periodic solutions of the dynamical systems, we offer an interactive computer algorithm based on finding the initial conditions corresponding to the periodic solutions with the possibility of interactive intervention and operational control of the computing process. We demonstrate the algorithm and various numerical examples of finding new and complex stable and unstable periodic solutions in strongly nonlinear dynamical systems with one and two degrees of freedom. We also consider the mutual influence of oscillations in multidimensional nonlinear dynamic systems.

1 Introduction

We consider nonlinear autonomous and nonautonomous dynamical systems of general form described by ordinary differential equations. Such dynamical systems correspond to forced, parametric, and self-excited vibration, where oscillations are generated by a combination of these modes. We also consider dynamical systems with several degrees of freedom. In such systems, depending on the values of the system parameters and external factors, diverse solutions exist. One of these solutions is a deterministic chaos [1]—chaotic behavior of solutions for the fully deterministic dynamic system parameters and external influences. However, stable periodic solutions of simple form (similar to the solutions of linear models of dynamic systems) are also possible. Typically, gradually changing parameters of the system lead

L. F. Petrov (✉)
Russian Plekhanov University of Economics, 36, Stremyanny per., Moscow, Russia
e-mail: lfp@mail.ru

National Research University Higher School of Economics, 20, Myasnitskaya Ulsitsa, Moscow, Russia

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_51

to a bifurcation, lose stability of solutions, leading to new solutions [2]. Other examples are known [3], while in the area of system parameters corresponding to a strange attractor, a stable periodic solution is found. We present a numerical algorithm for finding and analyzing the stability of a variety of periodic solutions of strongly nonlinear systems of ordinary differential equations. This algorithm allows us to investigate the evolution of periodic solutions when changing the parameters of the system. Such organization of work allows us to search for diverse stable and unstable periodic solutions for a variety of strongly nonlinear systems of ordinary differential equations including the area of deterministic chaos.

2 Problem Statement

Find the periodic solution of a strongly nonlinear dynamic system and check its stability.

2.1 Nonautonomous System: Finding of Periodic Solutions

For a system of differential equations of the form

$$\frac{dx_i}{dt} = X_i(x_1, x_2, ..., x_n, t),\ i = 1, 2, 3, ..., n, \tag{1}$$

$$X_i(x_1, x_2, ..., x_n, t) = X_i(x_1, x_2, ..., x_n, t + T)\ \text{ where } T \text{ is a known period} \tag{2}$$

find a kT-periodic solution of (1)

$$\varphi_i(t) = \varphi_i(t + kT)\ \ (k \text{ is a given number, } k = 1, 2, \ldots, K). \tag{3}$$

In this case, the problem is reduced to finding the initial conditions $Y_i = \varphi_i(0)$ $(i = 1, \ldots, n)$, corresponding to a periodic solution (3). To solve this problem, we have a system of nonlinear algebraic equations of the form

$$Y_i - x_i(kT) = 0,\ i = 1, 2, \ldots, n. \tag{4}$$

This system is solved using the Newton method.

2.2 Nonautonomous System: Stability of Periodic Solutions

To analyze the stability of the found kT-periodic solutions (3) the variational system is constructed. This is a linear system of ordinary differential equations with kT-periodic coefficients. This system is constructed by the analytically explicit functions X_i in system (1). Next, use the Floquet theorem, the monodromy matrix, and multipliers and calculate the characteristic Lyapunov exponents. These calculations allow us to determine the stability of the found kT-periodic solutions.

2.3 Autonomous Systems: Finding of Periodic Solutions

For a system differential equations of the form

$$\frac{dx_i}{dt} = X_i(x_1, x_2, \ldots, x_n),\ i = 1, 2, 3, \ldots, n, \tag{5}$$

find the T-periodic solution

$$\varphi_i(t) = \varphi_i(t + T) \tag{6}$$

and identify its period T. This problem is reduced to finding the initial conditions $Y_i = \varphi_i(0)$ $(i = 1, \ldots, n - 1)$, $Y_n = 0$, corresponding to the periodic solution, and period of solution T. To solve this problem, we have a system of nonlinear algebraic equations of the form

$$Y_i - x_i(T) = 0,\ i = 1, 2, \ldots, n - 1, \quad x_n(T) = 0. \tag{7}$$

This system can also be solved using the Newton method.

2.4 Autonomous Systems: Stability of Periodic Solutions

To analyze the stability of the found T-periodic solutions (6) of autonomous system (5) the variational system and determined multipliers are constructed.

For an autonomous system (5), we use the theorem of Andronov–Witt. If the periodic solution (6) $\varphi_i(t) \neq 0$, then the periodic solution is asymptotically stable if all the modules of multipliers $M_i < 1$ and with one $M_j = 1$.

In both variants 2.1 and 2.2 the dimension of systems algebraic Eqs. (4) and (7) is n, the algorithm for finding periodic solutions is iterative, the algorithm stability analysis is closed.

3 Realization of the Algorithm

Periodic solutions of these nonlinear dynamical systems can be complex, polyharmonic and have different periods for nonautonomous systems. The behavior of solutions by changing parameters is multivariate. In this situation, the interactive form of the algorithm can implement a search for various periodic solutions and analyze their evolution.

The scheme of algorithm is shown in Fig. 1. Interim results of the search of solutions are displayed in online mode. Users can interrupt the calculation process and realize a variety of control methods of the algorithm:

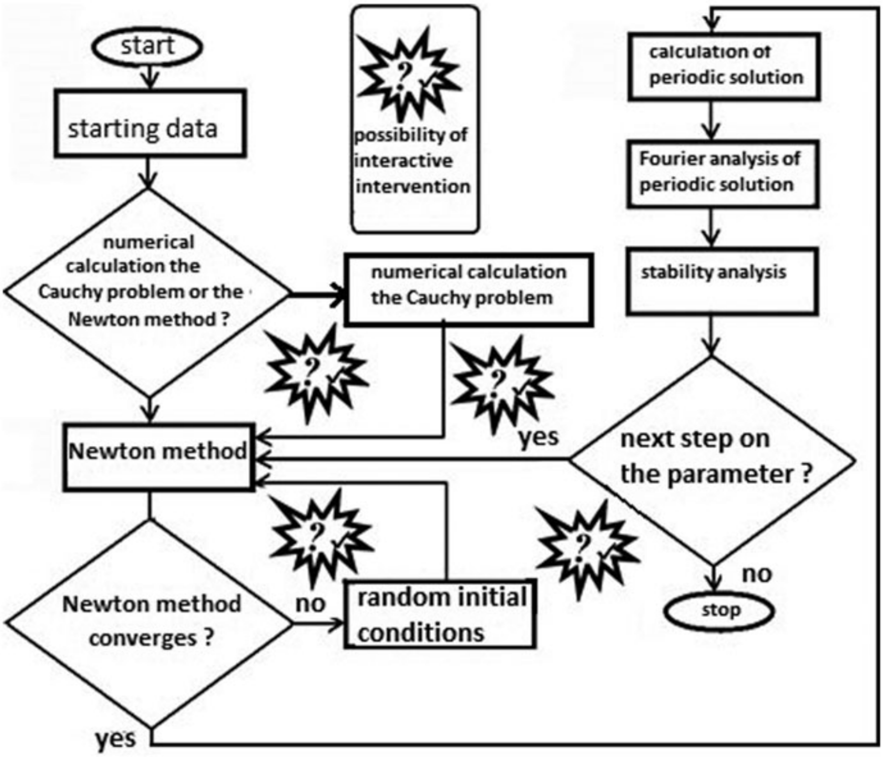


Fig. 1 Scheme of interactive computational algorithm search strategy of periodic solutions in essentially nonlinear dynamics

- Change searching strategy to move from the search, the initial conditions for periodic solutions to the Cauchy problem for a large time (a few periods of required solution), and vice versa
- Change any parameters of the system of differential equations
- Go back by steps to the parameters of the system
- Change the accuracy of all the used numerical methods
- Manage changes in the parameters of the system
- Operate a generator of random numbers for the initial conditions
- Manage the report generation and construction of phase trajectories, and so on

4 Examples of New Solutions

4.1 System Prototype

In the study of the system

$$\frac{d^2x(t)}{dt^2} - 10x + \frac{dx}{dt} + 100x^3 = W\cos 3.76t, \tag{8}$$

P. Holmes [3] found a strange attractor, and in the zone of attractor, he found a stable $5T$-periodic solution. This system is obtained in the study of transverse vibrations of the beam-hopping with one vibration mode.

4.2 Additional New Solutions

In addition to the periodic solutions found by Holmes in the area of strange attractor periodic solutions, we represent a new stable $3T$-periodic solution of the same system with $1.6 < W < 1.7$ (Fig. 2).

We also consider examples of the bifurcation of periodic solutions on the boundary of the strange attractor [4].

4.3 The Generalization of the Model (8) with Several Degrees of Freedom

$$\ddot{x}_j(t) + j^2(k^2j^2 + D)x_j(t) + x_j(t)\frac{j^2}{4}\left(\sum_{m=1}^{N} m^2x_m^2(t)\right) + \delta_1\dot{x}_j(t) = Q_j(t), \tag{9}$$

$$Q_j(t) = Q_j(t+T),\ j = 1, 2, \ldots, N \tag{10}$$

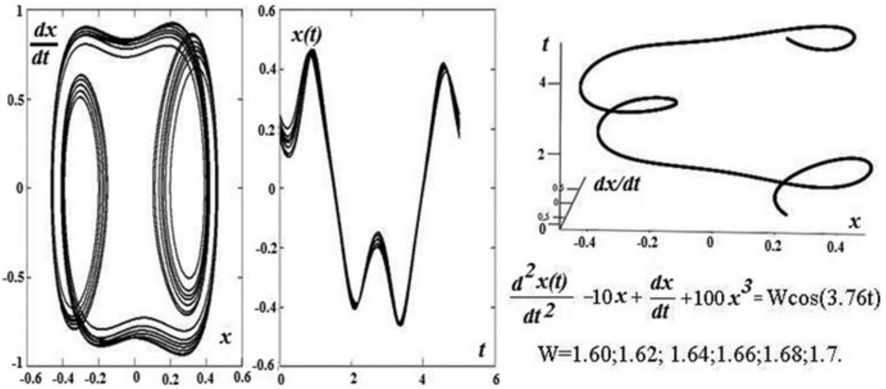


Fig. 2 Stable $3T$-periodic solution of system (8)

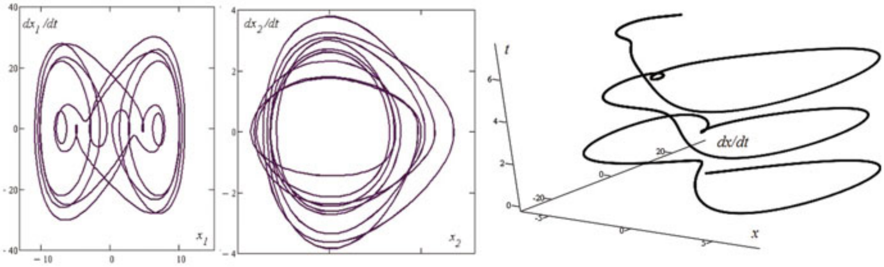


Fig. 3 The phase portraits of stable periodic solutions of the system (9) for $N = 2$

This system is obtained in the study of transverse vibrations of the beam-hopping with several forms of oscillations. The proposed interactive search algorithm for periodic solutions allows to find the complex dynamical modes in a multi-dimensional system (9). It is found that the interplay of various forms of oscillations changes the range of existence of deterministic chaos. Figure 3 shows the phase portraits of stable periodic solutions of the system (9) for $N = 2$.

5 Conclusion

The proposed interactive computational algorithm for the search of periodic solutions in an essentially nonlinear dynamics is effective in a variety of situations, including the field of deterministic chaos.

References

1. Petrov, L.: Control of dynamical regimes of systems with deterministic chaos. In: Abstracts of reports at II International conference "Optimization and applications," pp. 175–178. Petrovac, Montenegro (2011)
2. Petrov, L.: Nonlinear effects in economic dynamic models. Nonlinear Anal. Theory Methods Appl. **71**(12), e2366–e2371 (2009). (Elsevier)
3. Holmes, P.: A nonlinear oscillator with a strange attractor. Philos. Trans. R. Soc. A **292,** 419–448 (1979)
4. Feigenbaum, M.: Universal behaviour in nonlinear systems. Los Alamos Sci. **1,** 4–27 (1980)

Explosive Behavior in the Black–Derman–Toy Model

Dan Pirjol

Abstract We consider the simulation of the Black–Derman–Toy (BDT) model with log-normally distributed rates in the spot measure, in discrete time and with a continuous state variable. We note an explosive behavior in the Eurodollar futures convexity adjustment at a critical value of the volatility, which depends on maturity, rate tenor, and simulation time step size. In the limit of a very small time step, this singularity appears for any volatility, and reproduces the Hogan–Weintraub singularity, which is generic for short-rate interest rate models with lognormally distributed rates. The singular behavior arises from a region in the state space which is usually truncated off in finite difference and tree methods, or is very poorly sampled in Monte Carlo methods, and thus is not observed under usual simulation methods.

1 Introduction

The dynamics of interest rate models with log-normal volatility specification in continuous time is known to display singular behavior. This was first noticed in the context of the short-rate models by Hogan and Weintraub [7], who observed that Eurodollar futures convexity adjustments are infinite in the Dothan model [4], see also [1, 15]. This model is defined by the short-rate process

$$dr_t = \sigma r_t dW_t + a r_t dt, \tag{1}$$

with a, σ real constants. A similar divergence is observed in the Black–Karasinski model [7].

Similar singular behavior was observed in HJM models [6] with log-normal volatility structure. The forward rates in such models become infinite with nonzero

The information views and opinions set forth in this publication are those of the author, and are in no ways sponsored, endorsed or related to the business of J. P. Morgan.

D. Pirjol (✉)
J. P. Morgan, New York, NY 10172, USA
e-mail: dpirjol@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_52

probability [6], which implies zero prices for certain financial instruments which have only positive payoffs, such as zero coupon bonds.

In order to avoid these issues, modified versions of interest rate models with log-normal volatility specification have been proposed. For short-rate models, a modification of the Dothan model was proposed by Sandmann and Sondermann [15], which replaces the assumption of log-normality of the short rate r with that of log-normality of the effective annualized rate r_e defined as $r = \log(1+r_e)$. Assuming $dr_{e,t} = \sigma r_{e,t} dW_t$ leads to the short-rate model

$$dr_t = (1 - e^{-r_t})\sigma dW_t - \frac{1}{2}\sigma^2(1 - e^{-r_t})^2 dt. \tag{2}$$

For small rates $r_t \to 0$ the short rate is log-normally distributed, while for very large rates $r_t \gg 1$ the volatility specification becomes normal. For this model the divergence observed in the Dothan model disappears, and the Eurodollar future prices are finite [15] .

Similar modifications of the log-normal HJM model that avoid the singular behavior connected with the explosion of the forward instantaneous rate were proposed in [5]. These models replace the volatility specification of the log-normal HJM model with the assumption of log-normality of the forward effective rate $f_{\mathrm{e}}(t, T)$ with compounding period δ, defined by $f(t, T) = \frac{1}{\delta}\log(1+f_{\mathrm{e}}(t, T)\delta)$. This avoids the explosion of the forward rates [5]. The recognition of this fact led to the formulation of the modern log-normal Libor market model [3, 9, 10], which is defined in terms of the process for the set of finite tenor nonoverlapping simple forward rates $f(t, T, T+\delta)$.

In practice, most interest rate models are used in a discrete time implementation. We study here the emergence of the singular behavior in the discrete-time version of short-rate models with log-normally distributed rates. Using the example of the Black–Derman–Toy (BDT) model [2] with continuous state variable, we show the appearance of a sharp transition in the Eurodollar future convexity adjustment at a certain critical value of the volatility. At this point, this convexity adjustment has an explosive behavior, which thus introduces a limit on the applicability of the model.

2 Short-Rate Models

An important class of interest rate models used in financial practice is the class of the short-rate models. These models are defined by specifying a stochastic process for the short rate r_t of the form

$$dr_t = \sigma(t, r_t)dW_t + \mu(t, r_t)dt. \tag{3}$$

Some of the most popular models of this type are the Vasicek–Hull–White model, the CIR model and the Black–Karasinski models [1].

In practical applications, these models are usually simulated in a discrete time approach, on a tenor of dates $0 = t_0 < t_1 < \cdots < t_n$. We assume for simplicity that the simulation times are uniformly spaced $t_{i+1} - t_i = \tau$. The state variable r_t is either discretized (e.g., in finite difference or tree methods), or treated as a continuous variable (e.g., in Monte Carlo methods or analytical methods). We will consider here the second method, and will allow the state variable to be continuous and unbounded.

The price at time t_i of a European claim with payoff X at time t_j is given by the discrete time version of the fundamental pricing formula

$$V_i = \frac{1}{1 + L_i \tau} \mathbb{E}\left[\frac{X}{(1 + L_{i+1}\tau) \cdots (1 + L_{j-1}\tau)} |\mathcal{F}_i \right]. \tag{4}$$

The expectation value is taken in the so-called spot measure, with numeraire M_i given by the discrete-time analog of the money market account $M_i = \Pi_{k=0}^{i-1}(1 + L_k\tau)$. We will assume the use of the spot measure everywhere in the following.

The simplest claims are the zero coupon bonds, which correspond to $X = 1$. Denote $P_{i,j}$ the price of a zero coupon bond at time t_i paying 1 at time t_j. The Libor rate $L_{i,j}$ for the time period (t_i, t_j) is related to the zero coupon bond prices as $L_{i,j} = \frac{1}{t_j - t_i}\left(P_{i,j}^{-1} - 1\right)$. We denote the single-period rate as $L_{i,i+1} = L_i$.

2.1 BDT Model: Discrete-time, Continuous-State Implementation

The BDT model [2] is defined by the log-normal distributional assumption for the Libor rates L_i for the (t_i, t_{i+1}) period in the spot measure

$$L_i = \tilde{L}_i e^{\sigma_i x_i - \frac{1}{2}\sigma_i^2 t_i}, \quad i = 0, 1, \ldots, n-1, \tag{5}$$

where $\tilde{L}_i, \sigma_i$ are constants and $x_i \equiv x_{t_i}$ is a standard Brownian motion sampled at the simulation times t_i. The parameters $\tilde{L}_i, \sigma_i$ are calibrated such that the model reproduces a given initial yield curve $P_{0,i}$ and appropriate rate volatilities [8].

For the purpose of the simulation of the model, the zero coupon bonds $P_{i,j}(x_i)$ must be computed as functions of the stochastic driver x_i at the time t_i. The zero coupon bonds are given by (4) as conditional expectation values

$$P_{i,j}(x_i) = \frac{1}{1 + L_i\tau} \mathbb{E}\left[\frac{1}{\Pi_{k=i+1}^{j-1}(1 + L_k\tau)} |\mathcal{F}_{t_i} \right] = \frac{1}{1 + L_i\tau} \mathbb{E}\left[P_{i+1,j} |\mathcal{F}_{t_i} \right]. \tag{6}$$

This can be conveniently computed in a backward recursion in i. Introducing the functions $\pi_{i,j}(x_i)$ defined as

$$P_{i,j}(x_i) = \frac{1}{1 + L_i(x_i)\tau} \pi_{i,j}(x_i), \tag{7}$$

and the second equality in (6) gives the recursion relation

$$\pi_{i,j}(x_i) = \int_{-\infty}^{\infty} \frac{dx_{i+1}}{\sqrt{2\pi\tau}} e^{-\frac{1}{2\tau}(x_{i+1} - x_i)^2} \frac{\pi_{i+1,j}(x_{i+1})}{1 + L_{i+1}(x_{i+1})\tau} \tag{8}$$

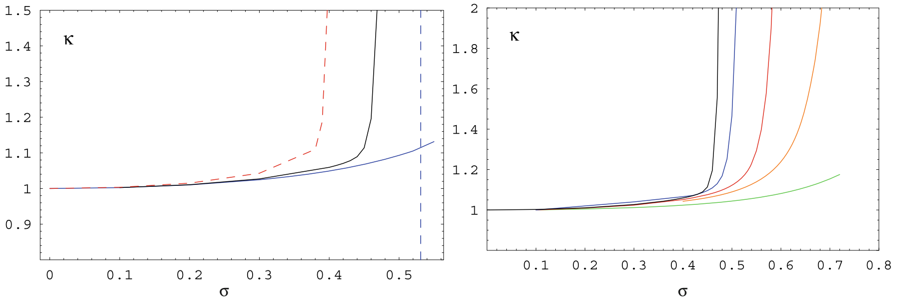


Fig. 1 *Left* plot: The Eurodollar future convexity adjustment in the BDT model for the Libor rate $L(T, T+\delta)$ set at $T = 5$ and paid at $T + \delta = 7.5$ as a function of the volatility σ in a simulation with time step $\tau = 0.25$. *Solid black curve*: continuous state variable, *solid blue curve*: binomial tree implementation. The *dashed curves* show analytical upper (*red*) and lower (*blue*) bounds [14]. *Right* plot: the same convexity adjustment vs σ for several time discretizations. The number of simulation time steps spanned by δ is $n_\delta = 2, 4, 6, 8, 10$ (from right to left). Both plots correspond to uniform forward rate $L_i^{\text{fwd}} = 5\%$ and $\tilde{L}_i = L_i^{\text{fwd}}$ (no calibration)

with initial condition $\pi_{j-1,j}(x) = 1$. The expectation value in (6) is expressed as an integral over the probability transition for the Brownian motion $p(x_{t_{i+1}}|x_{t_i})$, which is the heat kernel of the one-dimensional heat equation. The recursive integrations in (8) can be performed numerically using a precise method proposed in [13].

3 Eurodollar Future Pricing in the BDT Model

Consider the price V_t of a Eurodollar future on the rate $L(T, T+\delta)$. Assuming future settlement at discrete times t_i this is given by

$$V_t = 100(1 - \mathbb{E}[L(T, T+\delta)|\mathcal{F}_t]\delta). \tag{9}$$

The tenor of the Eurodollar future is δ (typically $3M$), and its delivery time is $T > t$.

We will compute the expectation of the Libor rate $L(T, T+\delta)$ in the spot measure; this can be parameterized in terms of a convexity adjustment κ defined as

$$\mathbb{E}\left[L(T, T+\delta)\right] = \kappa(T, \delta, n_\delta) L^{\text{fwd}}(T, T+\delta). \tag{10}$$

This expectation can be reduced to the calculation of the integral

$$\mathbb{E}\left[P_{i,j}^{-1}(x_i)\right] = \int_{-\infty}^{\infty} \frac{dx_i}{\sqrt{2\pi t_i}} e^{-\frac{1}{2t_i}x_i^2} (1 + L_i(x_i)\tau) \frac{1}{\pi_{i,j}(x_i)}. \tag{11}$$

We show in Fig. 1 typical results for the convexity adjustment κ vs. the volatility σ in the BDT model with uniform volatility $\sigma_i = \sigma$ and flat yield curve $L_i^{\text{fwd}} = 5\%$. The left plot shows κ both for the exact continuous state (solid black curve) and the usual binomial tree implementation of the BDT model (solid blue curve). They agree well for small σ, but diverge for larger volatility where the continuous state result has a

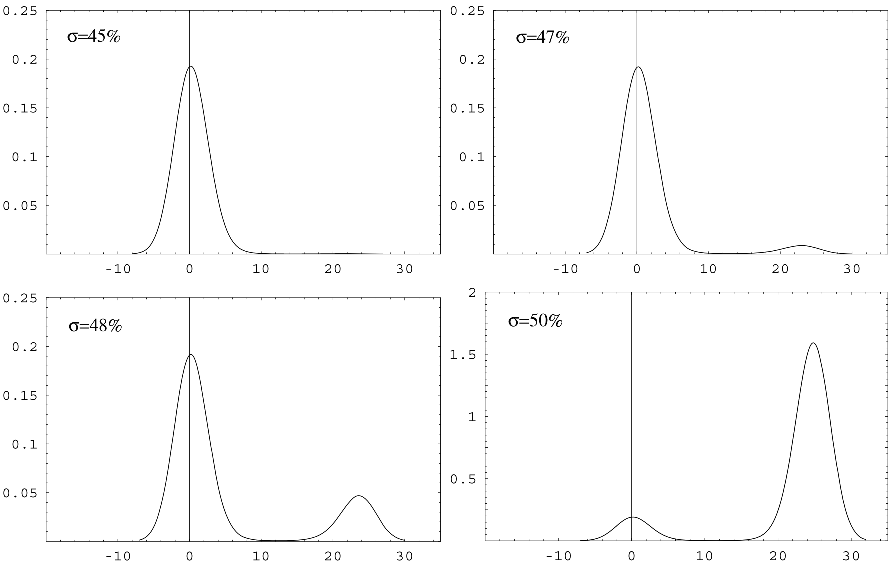


Fig. 2 The integrand in the expectation value (11) for the Eurodollar future convexity adjustment for the Libor rate $L(T, T+\delta)$ with $T=5, \delta=2.5$ for several values of the volatility σ around the critical value $\sigma_{\mathrm{cr}}=47\,\%$. The simulation has quarterly time steps $\tau=0.25$ and $L_i^{\mathrm{fwd}}=5\,\%$

sharp increase at a critical value $\sigma_{\mathrm{cr}} \sim 47\,\%$. This shows that the usual simulation methods can miss the correct explosive behavior of the model. The emergence of the sharp transition as the number of simulation time steps $n_\delta=\delta/\tau$ spanned by the Libor tenor increases is seen in the right plot of Fig. 1.

In order to investigate the origin of this singularity, we show in Fig. 2 the integrand of the expectation value (11) for several values of the volatility around the transition point $\sigma_{\mathrm{cr}} \simeq 47\,\%$. For small volatility below the critical value, the integrand is peaked around the origin with width $\sim\sqrt{t_i} \simeq 2.24$. This is the region covered well in usual tree and finite difference implementations. Around the critical volatility the integrand develops a second peak at a very large value of $x \sim 24$, which rapidly increases in size as the volatility crosses the critical value. Above σ_{cr} the integral (11) is dominated by the second peak. The latter lies at about 10 standard deviations of x_i from origin, in a region which is assumed to be unimportant in practice, and is truncated off in usual implementations of the model. Accounting for the contribution of the secondary peak at $x \sim 24$ requires a very high precision of the simulation; at this point, the denominator in (11) is equal to $\pi_{i,j}(x) \sim 10^{-21}$ (for $\sigma=47\,\%$).

A similar phenomenon was observed in models with log-normally distributed rates in the terminal measure [11, 12]. In these models, certain expectation values and convexity adjustments were shown to have a sharp transition and explosive behavior as the function of volatility. As in the case considered here, the effect is due to a contribution to the expectation values from a region in the state variable that is assumed to be unimportant and is truncated off in usual simulation methods.

4 Conclusions

We report the results of an investigation of the BDT model with continuous state variable in discrete time. The main conclusions of this study are:

- In discrete time, the Eurodollar convexity adjustment for rates spanning sufficiently many simulation steps has a singular behavior at a finite but nonzero value of the volatility. At this point, the convexity adjustment explodes to unphysically large values, which limits the applicability of the model.
- The singular behavior arises from a region in the state space which is usually truncated off in finite difference and tree methods, or very poorly sampled in Monte Carlo methods, and thus is not observed under usual simulation methods.
- The critical volatility decreases with the time step and approaches zero in the continuous time limit, in agreement with the Hogan–Weintraub result [7, 15]. This is suggested by numerical simulation (e.g., in Fig. 1 (right panel)) and is confirmed by an analytical lower bound on κ [14].

References

1. Andersen, L., Piterbarg, V.: Interest Rate Modeling. Atlantic Financial Press London, UK (2010)
2. Black, F., Derman, E., Toy, W.: A one-factor model of interest rates and its application to treasury bond options. Financ. Anal. J. **46**, 24–32 (1990)
3. Brace, A., Gatarek, D., Musiela, M.: The market model of interest rate dynamics. Math. Financ. **7,** 127–154 (1997)
4. Dothan, L.: On the term structure of interest rates. J. Financ. Econ. **6,** 59–69 (1978)
5. Goldys, B., Musiela, M., Sondermann, D.: Log-normality of rates and term structure models. Stoch. Anal. Appl. **18**(3):375–396 (2000)
6. Heath, D., Jarrow, R., Morton, A.J.: Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. Econometrica **60,** 77–105 (1992)
7. Hogan, M., Weintraub, K.: The lognormal interest rate model and eurodollar futures. Citibank (1993)
8. Jamshidian, F.: Forward induction and construction of yield curve diffusion models. J. Fixed Income **1,** 62–74 (1991)
9. Jamshidian, F.: Libor and swap market models and measures. Financ. Stoch. **1,** 293–330 (1997)
10. Miltersen, K., Sandmann, K., Sondermann, D.: Closed form solutions for term structure derivatives with log-normal interest rates. J. Financ. **52**(1):409–430 (1997)
11. Pirjol, D.: Phase transition in a log-normal Markov functional model. J. Math. Phys. **52,** 013301 (2011)
12. Pirjol, D.: Explosive behavior in a log-normal interest rate model. Int. J. Theor. Appl. Financ. **16,** 1350023 (2013)
13. Pirjol, D.: The logistic-normal integral and its generalizations. J. Comput. Appl. Math. **237,** 460–469 (2013)
14. Pirjol, D.: Hogan-Weintraub singularity and explosive behaviour in the Black-Derman-Toy model, to appear in Quantitative Finance
15. Sandmann, K., Sondermann, D.: A note on the stability of lognormal interest rate models and the pricing of Eurodollar futures. Math. Financ. **7**(2):119–125 (1997)

Exploiting Block Triangular form for Solving DAEs: Reducing the Number of Initial Values

J. Pryce, N. Nedialkov, G. Tan and R. McKenzie

Abstract The authors have written two codes to solve differential algebraic equations (DAEs) by structural analysis (SA). The first is written in C++ (DAETS) and deals with the solution of DAE initial value problems, using SA. Upon seeing how informative the SA could be the authors wrote DAESA (in MATLAB) to do only the structural analysis. These codes rely on exploiting the block triagular form (BTF) of a DAE, this chapter explains how.

1 Overview of the Structural Analysis (SA) Method

Both DAETS and DAESA handle a DAE in n state variables $x_j(t)$, $j = 1, \ldots, n$, of the general (possibly nonlinear) form

$$f_i(t,\ \text{the } x_j \text{ and derivatives of them}) = 0, \quad i = 1, \ldots, n,$$

which includes the case of a fully implicit or purely algebraic system. The numerical solution scheme used in DAETS is via Taylor series, in steps over a range, using automatic differentiation, analogous to a Taylor series method for ODEs.

The method starts by forming the $n \times n$ signature matrix $\Sigma = (\sigma_{ij})$, where

$$\sigma_{ij} = \begin{cases} \text{highest order of derivative to which } x_j \text{ occurs in } f_i \\ -\infty \text{ if it does not occur.} \end{cases} \tag{1}$$

J. Pryce (✉) · R. McKenzie
Cardiff University, Cardiff, UK
e-mail: smajdp1@cardiff.ac.uk

N. Nedialkov · G. Tan
McMaster University, Hamilton, ON, Canada
e-mail: nedialk@mcmaster.ca

G. Tan
e-mail: tang4@mcmaster.ca

R. McKenzie
e-mail: mckenzier1@cardiff.ac.uk

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_53

A highest value transversal (HVT) is found, which comprises n finite entries, one in each row and column of Σ, such that the total of these entries is maximised. We assume the problem is structurally well-posed, meaning that such an HVT exists. Valid non-negative integer valued offset vectors $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{d} = (d_1, \ldots, d_n)$ are found, where valid means

$$d_j - c_i \geq \sigma_{ij} \quad \text{for all } i, j, \text{ with equality on a HVT,} \tag{2}$$

normalised by the constraint $\min_i c_i = 0$. There are unique element-wise smallest vectors $\mathbf{c}$, $\mathbf{d}$, which we call the canonical offsets; these are used from now on. However, any choice of valid offsets specifies a solution scheme by which to find Taylor coefficients in batches. Namely for stage $k = k^*, k^* + 1, \ldots$ where $k^* = -\max_j d_j \leq 0$, solve the equations

$$f_i^{(k+c_i)} = 0 \quad \forall i \text{ such that } k + c_i \geq 0 \tag{3}$$

for the variables

$$x_j^{(k+d_j)} \quad \forall j \text{ such that } k + d_j \geq 0. \tag{4}$$

Consider the simple pendulum DAE, in variables $x(t), y(t), \lambda(t)$ and parameters length L and gravity G. We find its Σ matrix, HVT (two, one marked by $\bullet$ the other by $\circ$) and offsets

$$\begin{aligned} 0 = A &= x'' + x\lambda \\ 0 = B &= y'' + y\lambda - G \\ 0 = C &= x^2 + y^2 - L^2 \end{aligned} \qquad \Sigma = \begin{array}{c|ccc|c} & x & y & \lambda & c_i \\ \hline A & 2^\bullet & -\infty & 0^\circ & 0 \\ B & -\infty & 2^\circ & 0^\bullet & 0 \\ C & 0^\circ & 0^\bullet & -\infty & 2 \\ \hline d_j & 2 & 2 & 0 & \end{array}$$

This specifies a solution scheme of the form

Stage k	Solve	For ‘batch’	Kind
-2	$0 = C = x^2 + y^2 - L^2$	x, y	1by2 nonlinear
-1	$0 = C' = 2xx' + 2yy'$	x', y'	1by2 linear
0	$0 = A, B, C''$	x'', y'', λ	3by 3linear
1	$0 = A', B', C'''$	x''', y''', λ'	3by3 linear

(5)

and so on for later stages. At each stage, treat items found previously as ‘known’.

A key object is the $n \times n$ system Jacobian matrix $\mathbf{J}$, with entries

$$\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j^{(d_j - c_i)}} = \begin{cases} \dfrac{\partial f_i}{\partial x_j^{(\sigma_{ij})}} & \text{if } d_j - c_i = \sigma_{ij} \\ 0 & \text{otherwise, including where } \sigma_{ij} = -\infty \end{cases}$$

The solution scheme succeeds [3] iff $\mathbf{J}$ is non-singular, for example, for the pendulum,

$$\mathbf{J} = \begin{bmatrix} \partial A/\partial x'' & 0 & \partial A/\partial\lambda \\ 0 & \partial B/\partial y'' & \partial B/\partial\lambda \\ \partial C/\partial x & \partial C/\partial y & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 2x & 2y & 0 \end{bmatrix}$$

is non-singular, since $\det(\mathbf{J}) = -2(x^2 + y^2) = -2L^2 \neq 0$ at a consistent point.

2 An IV Paradox and Its Explanation

IVs, Naive Version

Solution scheme (3), (4) gives a simple recipe, in terms of the offsets, for what initial values (IVs) must be provided: namely these comprise all $x_j^{(r)}$ such that

$$\begin{cases} 0 \le r < d_j, & \text{if the DAE is quasilinear (see below), so } \sum_j d_j \text{values in all;} \\ 0 \le r \le d_j, & \text{otherwise, so } n + \sum_j d_j \text{values in all.} \end{cases} \tag{6}$$

For example, the simple pendulum is quasilinear with $\mathbf{d} = (2, 2, 0)$, so the recipe is: in scheme (5), give IVs for x, x'; y, y'. *Note:* we call them IVs but they are really a set of *trial values* from which a nearby initial consistent point can always be computed in a numerically stable way. Since this DAE has 2 degrees of freedom (DOF) one could specify such a point choosing just two of these values, say x, x', instead of four; but any such choice is numerically unstable for some initial position of the pendulum.

IVs, Exploiting DAE Structure When applied to DAEs having structure, the simple recipe (6) leads to paradoxes. Namely, if one subsystem of the DAE drives another but is itself undriven, (6) can make the driving subsytem need more IVs than it would, were it stand-alone. An example is the coupled two pendula system

$$\begin{aligned} 0 &= A = x'' + x\lambda, \\ 0 &= B = y'' + y\lambda - G, \\ 0 &= C = x^2 + y^2 - L^2, \\ 0 &= D = u'' + u\mu, \\ 0 &= E = v'' + v\mu - G, \\ 0 &= F = u^2 + v^2 - (L + cx')^2\,. \\ & \quad c \text{ is a constant}\,. \end{aligned} \qquad \Sigma = \begin{array}{c|ccc|ccc|c} & x & y & \lambda & u & v & \mu & c_i \\ \hline A & 2 & & 0 & & & & 1 \\ B & & 2 & 0 & & & & 1 \\ C & 0 & 0 & & & & & 3 \\ \hline D & & & & 2 & & 0 & 0 \\ E & & & & & 2 & 0 & 0 \\ F & 1 & & & 0 & 0 & & 2 \\ \hline d_j & 3 & 3 & 1 & 2 & 2 & 0 & \end{array}, \tag{7}$$

where a blank in the Σ means $-\infty$. The first three equations model one pendulum, and the second three model a second (coupled) pendulum. Pendulum 1 drives pendulum 2 in that x' appears in equation F (horizontal velocity of pendulum 1 affects length of pendulum 2), see entry in position (F, x) of Σ; but there's no reverse influence.

Yet, pendulum 1's offsets have increased by 1 while pendulum 2's are unchanged. Hence, pendulum 1 now needs IVs for x'', y'' and λ, which as a stand-alone system it did not. Clearly something is wrong here.

To clarify what is going on, consider Table 1, which lists the stages in the uncoupled and coupled system solution processes.

Uncoupled pendula	
k	find
-2	$x,y,\quad u,v$
-1	$x',y',\quad u',v'$
0	$x'',y'',\lambda,\quad u'',v'',\mu$

(a)

Coupled pendula	
k	find
-3	x,y
-2	$x',y',\quad u,v$
-1	$x'',y'',\lambda,\quad u',v'$
0	$x''',y''',\lambda',\quad u'',v'',\mu$

(b)

Table1 Stages in solving the uncoupled and coupled 2-pendula systems.

Uncoupled pendula		
k	find in parallel	
-2	x,y	u,v
-1	x',y'	u',v'
0	x'',y'',λ	u'',v'',μ

(a)

Coupled pendula			
local k^* $= k+1$	find	global k	*then* find
-2	x,y	-3	
-1	x',y'	-2	u,v
0	x'',y'',λ	-1	u',v'
1	x''',y''',λ'	0	u'',v'',μ

(b)

Table2 Explanation in terms of local stage counter k^* for first pendulum.

With coupling, you can't find u, v at same time as x, y, because u, v satisfy $0 = F = u^2 + v^2 - (L + cx')^2$, which uses x' which hasn't been found yet. And so on. The raised offsets say, in effect, 'Shift pendulum 1 a stage earlier, so its derivatives are ready when pendulum 2 needs them'. Their apparent effect of increasing the number of IVs needed is mistaken, and due to (6) not telling the whole story.

The explanation comes from considering pendulum 1 to have a *local stage counter* $k^* = k + 1$. (Pendulum 2 also has one, but it is the same as the global counter k.) Introducing k^* into Table 1 gives Table 2.

In the coupled system *each global stage* solves for pendulum 1 data first and then uses this to solve for pendulum 2 data. Relative to its *local* stages, pendulum 1 has local offsets $(\widehat{\mathbf{c}}, \widehat{\mathbf{d}}) = (0, 0, 2; 2, 2, 0)$, the same as when it is stand-alone. And it is clear from the solution scheme in Table 2b that pendulum 1 requires the same IVs x, y, x', y' as when it is stand-alone—which is as it should be.

3 The Benefits of BTF

The paradox in the previous section arose because the DAE had a *nontrivial block triangular form (BTF)*, and the explanation came from recognising this.
A BTF is a property of a *sparsity pattern*, in this case a subset S of $\{1, 2, \ldots, n\}^2$, the $n \times n$ positions in the signature matrix or Jacobian. Write a × in the positions (i, j) that belong to S, and leave the others blank.

If we can permute rows and columns so S looks like the example below, then we have put S in (upper) block triangular form.

$$\left[\begin{array}{cc|ccc|ccc}
\times & \times & & \times & & \times & & \times \\
\times & \times & & \times & \times & & & \\
\hline
 & & \times & \times & & & \times & \\
 & & \times & \times & \times & & & \\
 & & \times & & \times & \times & & \times \\
\hline
 & & & & & & \times & \times \\
 & & & & & \times & \times & \\
 & & & & & \times & & \times
\end{array}\right],$$

Namely, there are square diagonal blocks that are themselves irreducible (cannot be split into a finer BTF) and the below-diagonal blocks are empty. Such a BTF can always be found if S is structurally non-singular, i.e. contains some transversal; and it is unique up to ordering of the diagonal blocks [1].

In the DAE structural analysis context we have two choices of sparsity pattern to use. A natural one is the sparsity pattern of Σ:

$$S = \{(i, j) \mid \sigma_{ij} > -\infty\}$$

We call the BTF based on this sparsity pattern the *coarse BTF*. A more informative BTF is found by using the sparsity pattern of the Jacobian $\mathbf{J}$:

$$S_0 = S_0(\mathbf{c}, \mathbf{d}) = \{(i, j) | d_j - c_i = \sigma_{ij}\}$$

We call the resulting BTF the *fine BTF* since $S_0 \subseteq S$ and it usually gives a strict refinement of the BTF that S generates. Though S_0 depends on the $(\mathbf{c}, \mathbf{d})$ chosen, the resulting set of blocks is independent of $(\mathbf{c}, \mathbf{d})$ up to possible reordering [2].

Each diagonal block of the BTF defines a subsystem of the DAE: its equations (rows) and variables (columns) form a free-standing DAE, if one counts any other variables that occur in these equations as external driving functions.

Let there be p blocks of sizes $N_1, \ldots, N_p$ summing to n. As we are using upper BTF, each block depends only on those below it, the bottom block being independent of all others. It can be proved that for the *fine* BTF, there is a well-defined notion that the ℓth block has a *local stage counter*

$$k_\ell = k + K_\ell, \quad \ell = -1, \ldots, p,$$

where k is the global stage counter and K_ℓ is an integer ≥ 0, the *lead time* of that block. The *local offsets* given by

$$\widehat{c_i} = c_i - K_\ell, \qquad \widehat{d_j} = d_j - K_\ell$$

are thus the offsets of the ℓth block as a free-standing DAE in the sense described above.

As an example, the coupled two pendula DAE (7) has the following fine BTF:

$$\Sigma = \begin{array}{c|ccc|ccc|cc} & u & v & \mu & x & y & \lambda & c_i & \widehat{c_i} \\ \hline F & 0 & 0 & & 1 & & & 2 & 2 \\ E & & 2 & 0 & & & & 0 & 0 \\ D & 2 & & 0 & & & & 0 & 0 \\ \hline C & & & & 0 & 0 & & 3 & 2 \\ B & & & & & 2 & 0 & 1 & 0 \\ A & & & & 2 & & 0 & 1 & 0 \\ \hline d_j & 2 & 2 & 0 & 3 & 3 & 1 & & \\ \widehat{d_j} & 2 & 2 & 0 & 2 & 2 & 0 & & \end{array}, \qquad \mathbf{J} = \begin{array}{c|ccc|ccc|} & u & v & \mu & x & y & \lambda \\ \hline F & 2u & 2v & & \xi & & \\ E & & 1 & v & & & \\ D & 1 & & u & & & \\ \hline C & & & & 2x & 2y & \\ B & & & & & 1 & y \\ A & & & & 1 & & x \\ \hline \end{array}$$

where $\xi = -2c(L+cx')$. Hence we have a lead time $K_1 = 0$ for the first block and $K_2 = 1$ for the 2nd block.

4 Initial Values Revisited

Initial (or trial) values are numbers a user must supply, to specify an initial consistent point of the DAE from which to propagate a numerical solution. Of course one would like to demand as few IVs as possible. The BTF theory outlined above (see [2] for more detail) shows the IVs needed are determined, not by the naive recipe (6) using global offsets, but by the corresponding formula using local offsets $\widehat{d_j}$. Namely within a block they comprise all $x_j^{(r)}$ such that:

$$\begin{cases} 0 \leq r < \widehat{d_j}, & \text{if the block is quasilinear (see below);} \\ 0 \leq r \leq \widehat{d_j}, & \text{otherwise;} \end{cases} \tag{9}$$

where quasilinear (QL) means the equations in the block are jointly linear in their leading derivatives, $x_j^{(d_j - c_i)}$. With blocks of size N_ℓ and lead times K_ℓ, $\ell = 1, \ldots, p$, this needs $\sum_{\ell=1}^{p} N_\ell K_\ell$ fewer IVs than recipe (6). This is before considering an important effect: experience suggests that small blocks are far more likely to be QL than are large ones. As a result, BTF and QL analysis work together to reduce the number of IVs needed.

5 Conclusions and Future Work

Structural analysis based on BTF promises to reduce the work of solving a DAE IVP numerically, especially by simplifying the linear algebra involved. As yet, DAETS does not use such methods, whereas some other simulation tools based on DAE models do use some form of SA to reduce work.

The coarse rather than fine, BTF can be exploited to solve IVPs in parallel by pipelining the solution process block-wise: this also deserves study.

We believe that our theory is the most powerful available, for doing this sort of analysis, but other tools are currently ahead of DAETS in putting SA into practice. Thus, we wish to study other simulation tools' use of SA in numerical solution.

References

1. Pothen, A., Fan, C.J.: Computing the block triangular form of a sparse matrix. ACM Trans. Math. Softw. **16**(4):303–324 (1990). http://doi.acm.org/10.1145/98267.98287
2. Pryce, J., Nedialkov, N.S., Tan, G.: DAESA—a Matlab tool for structural analysis of DAEs: theory. ACM Transactions on Mathematical Software, Vol. 41, No. 2, Article 9, January 2015 (20 pp)
3. Pryce, J.D.: A simple structural analysis method for DAEs. BIT **41**(2):364–394 (2001)

Analysis and Visualization of a Many-Objective Optimization Landscape Design Problem

Luis A. Rivera-Zamarripa, Steven A. Roberts and Nareli Cruz-Cortés

Abstract A general methodology to analyze the solution of a many-objective optimization problem (MOOP) to landscape design is presented. The landscape design problem consists on assigning different types of land use to specific areas identified as *candidate sites* to be changed. Some local and global ecological criteria are considered. In order to gain some clarity during the analysis of the solutions that conform to an optimal set some clustering strategies are usually utilized. In this chapter, we use a simple strategy called favour ranking to place similar solutions together. Then, the solutions are visualized using a state-of-the-art technique.

1 Introduction

The multiobjective optimization problems have been widely studied for the case of two or three objectives during the last decades. However, the case of four or more objectives has been heavily explored only during the last few years, they are called many-objective optimization problems (MOOPs). In these cases, some important problems are usually faced, for instance, the approximation of the optimal solution set (Pareto optimal solution set), visualization, and decision-making. According to the Pareto optimal definition for MOOP, all the solutions in the Pareto set (incomparable solutions) are considered as the same rank. However, for optimization methods such as the evolutionary algorithms, most of the time it is necessary to make a

Nareli Cruz-Cortés acknowledges support from CONACYT through the project 132073.

L. A. Rivera-Zamarripa (✉) · N. Cruz-Cortés
Centro de Investigación en Computación del Instituto Politécnico Nacional, México City, México
e-mail: lrivera_a13@sagitario.cic.ipn.mx

S. A. Roberts
Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: sroberts@wlu.ca

N. Cruz-Cortés
e-mail: nareli@cic.ipn.mx

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_54

distinction between them. Several strategies have been proposed to address this problem (for a comprehensive survey see [3]). For example, in [2], the authors used the relation *favour* to distinguish between solutions. The relation *favour* makes comparable solutions which are incomparable under Pareto dominance.

On the other hand, plotting a set of Pareto solutions for four or more objectives is a difficult task, therefore analyzing and choosing a solution from that set are difficult tasks too. Some published works have suggested the use of clustering techniques to place similar solutions together. These techniques usually give low clarity and are time consuming. Further, in some cases, the motivation to place solutions together is unclear. The work published in [9] suggested the usage of heatmaps as a very convenient alternative to visualize Pareto Front (PF^*) with high dimensionality. Mainly because they do not lose any information, and their efficiency is largely unaffected by the problem dimensionality. However, some interpretation problems arise if the solutions and objectives are placed in an arbitrary order [9]. The Pareto optimality definition allows the existence of solutions where a small improvement in only one objective can occur, while a large deterioration is present in some other objectives. The relation *favour* fixes, to certain degree, this problem also by considering the quantity of objectives where a solution is better than other. Further, giving a particular hierarchy to the solutions could help the decision-maker on his/her task by providing a clue on how the solutions are related. Therefore, when ordering according to favour, some optimal solutions would be considered better than others. Additionally, this ordering will place together those solutions that behave similar, e.g., solutions with a large number of functions with higher values, and vice versa. In this chapter, we propose using the relation called *favour* to impose an ordering to the solutions. The results can then be visualized using a heatmap.

In some problem domains such as Greenland System design, the statement of design problems using multiple objectives has arisen. It allows the trial and incorporation of different configurations based on ecological principles. Also, the analysis of qualitative and quantitative regional landscape structures in land-use planning [5]. The work published in [5] presented the statement of eight objective functions to find the best configuration for some portion land considered as sites candidates to be changed. A new statement for the problem is presented in [6] where the objectives were redesigned, and some of them established as constraints. In this chapter, we present the analysis of this multiobjective landscape design problem.

2 Multiobjective Optimization Problems

2.1 Basic Definitions

The task of a MOOP is to find the vector of variables $x^* = [x_1^*, x_2^*, x_3^*, \ldots, x_n^*]^T$ such that it optimizes the k objective functions $f(x) = [f_1(x), f_2(x), \ldots, f_k(x)]^T$ subject to inequality and equality constraints. The solutions that fulfill the constraints define the set of feasible solutions Ω. The vector variables x can be continuous, discrete or a combination of both kind of values.

Pareto dominance definition A solution $u = [u_1, u_2, \ldots, u_k]^T \in \Omega$ dominates a solution $v = [v_1, v_2, \ldots, v_k]^T \in \Omega$ (denoted as $u \preceq v$), if and only if, u is partially less than v, that is, $\forall i \in \{1, 2, \ldots, k\} : u_i \leq v_i \wedge \exists i \in \{1, 2, \ldots, k\} : u_i < v_i$.

Optimal Pareto set Given a MOOP $f(x)$, the Pareto optimal set (P^*) is defined as: $P^* = \{x \in \Omega \mid \neg\exists x' \in \Omega : f(x') \preceq f(x)\}$.

Pareto optimal front The PF^* is defined as: $PF^* = \{u = F(x) \mid x \in P^*\}$.

Then, the MOOP solution is comprised of set of compromise solutions (instead of only one point), which is precisely the optimal Pareto set. The difficulties faced when dealing with MOOP are mainly: high computational cost, poor scalability of most existing evolutionary algorithms, and difficulty to visualize the solutions. To deal with these high-dimension MOOP, the next steps should be followed:

- To approximate the optimal PF^*
- To remove the redundant objectives
- To apply some clustering technique to the solutions
- To visualize the solutions to choose one

2.2 Objective Reduction

To deal with the scalability problem, some authors have proposed reducing the number of objective functions, in which case, most of the mentioned difficulties will be diminished or eliminated.

Only few published works related to objective reduction in MOOP can be found [4, 8]. Some approaches assume the existence of some redundant objectives that can be removed from the objective set without affecting the Pareto relations [1]. On the other hand, other authors [7] consider that the objectives elimination must guarantee the preservation of the global correlation structure (instead of the Pareto relations). In [7], it is proposed an algorithm based on principal component analysis to denoise and reduce the number of objectives.

2.3 Visualization

Some techniques for visualization have been applied to MOOP fronts such as, scatter plots, pairwise coordinate plots, neuroscale, parallel coordinate plots, among others. According to the study presented in [9], *heatmaps* seem to be a convenient option.

3 Many-Objective Landscape Design Problem

In this chapter, we consider a MOOP combinatorial problem defined in [5, 6] where there are some land areas considered as candidates to be changed. The goal is to define, for each candidate site, which would be the best option to be changed considering one out of the three next values: natural, agricultural and urban. Each land configuration is evaluated according to seven objective functions that consider local as well as global ecological criteria. These objective functions are summarized next:

- F1: It rewards solutions with a few large patches of natural vegetation (core areas)
- F2: This function favors connected structures of natural features across the landscape by maximizing the vegetated corridors between core areas
- F3: It rewards the formation of stepping stones of natural vegetation between large natural areas
- F4: It maximizes the number of natural–urban neighbors in urban areas
- F5: It maximizes the number of agricultural–agricultural neighbors
- F6: This function favors agricultural areas considering soil capability
- F7: It maximizes the number of urban–urban neighbors

Further details related to these objective functions, and their ecological support can be found in [5, 6]. The case of study is the area located within the Greater Toronto Area, Ontario, Canada, where eight candidate sites were found.

Obtaining the optimal PF^* Taking advantage of the relatively small search space, the optimal PF^* was obtained by enumeration. The whole search space is composed by $3^8 = 6561$ possible solutions. From there, it was found that the optimal PF^* is conformed by $PF^* = 47$ solutions.

Objective reduction Based on the idea presented in [1] an objective reduction was applied. So, each objective was removed at a time and the Pareto relations computed and compared against the original ones. If the Pareto relations changed then, that objective is considered as essential. The next was found:

- Redundant objectives: F2 and F4
- Essential objectives: F1, F3, F5, F6, and F7.

3.1 Visualization Using a Ranking Based on Favour

In this chapter, we use the relation called *favour* to order the solutions to be shown in a heatmap. The relation *favour* is able to compare solutions which are incomparable using the relation *dominance* [2] (assuming the minimization case). The relation *favour* is defined as follows: let be two nondominated solutions u and v, defined in the objective space with k objective functions $u = [u_1, \ldots, u_k]$ and $v = [v_1, \ldots, v_k]$. It is said that u is better than v, $(u <_f v)$, if it is higher the number of objectives in which u is less than v, or more formally:

$$u <_f v \Leftrightarrow |\{i : f_i(u) < f_i(v), 1 \leq i \leq k\}| > |\{j : f_j(v) < f_j(u), 1 \leq j \leq k\}|.$$

For each solution u its score $s(u)$ is computed as the number of times that u is better than all other solutions in the front $(u <_f m)$, where $m \in PF$ and PF is the current PF^* approximation. The process of determining the score s has a complexity $O(kn^2)$,

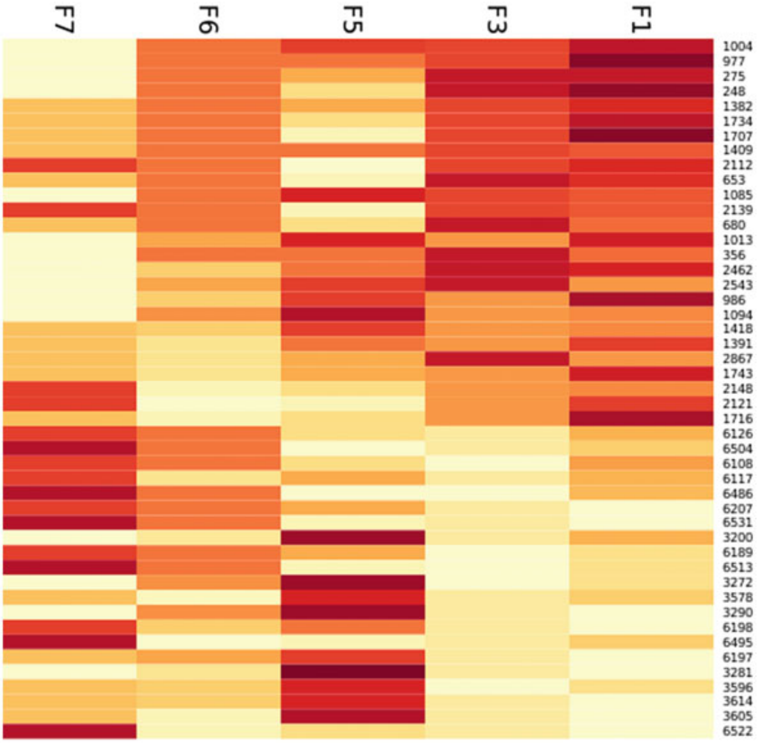


Fig. 1 Heatmap for the studied landscape design problem solution (*color online*)

where n is the number of individuals conforming the PF^*, and k is the number of objectives.

The solutions in PF are ordered according to their score $s(\cdot)$. The solutions with highest scores are placed at the top of the list, and vice versa. Notice that some ties can be found, that is, two solutions u and v could have the same score ($s(u) = s(v)$).

Ties are solved by considering not only the number of times that the solutions are better than other solutions, but the number of objectives in which it wins too. In this case, we say that u is better than v if the next condition is satisfied:

$$\{i : f_i(u) < f_i(m), u <_f m, \forall m \in PF\} > \{j : f_j(v) < f_j(m), v <_f m, \forall m \in PF\},$$

where $1 \leq i, j \leq k$.

The ordered solutions are then presented in a heatmap. For the studied problem, the solutions (without constraints) are presented next. The heatmap shows the solutions in the rows and the objectives at the top (see Fig. 1). The colors inside the heatmap represent the number of times that the solution wins to other solutions for the corresponding objective. Thus, the values for all the objectives are in the same scale.

4 Conclusions

Defining a design problem as multiobjective gives the possibility of exploring and analysing new configurations, without giving any objective preference before the optimization process. In such case, the decision-maker can consider all the possibilities. The application of some strategy to order the nondominated solutions gives a clue to the decision-maker to make an easier process. Ordering the solutions according to the relation *favour*, introduces some bias to the decision-making process. Actually, in this case, it is considered that the best solutions are those with highest values for more objective functions, which implies that they have a very low value in at least one objective. This clustering process does not need to define any similarity functions or user-defined parameters. Instead, it only compares each solution against the others to count the number of times that it is better in terms of the relation *favour*. Clearly, it is a simpler process than the one presented in [9] where a similarity matrix is conformed by obtaining its eigenvalues.

References

1. Brockhoff, D., Zitzler, E.: Are all objectives necessary: on dimensionality reduction in evolutionary multiobjective optimization. In: Parallel Problem Solving from Nature IX, pp. 533–542, Iceland, September 9–13 (2006)
2. Drechsler, N., Drechsler, R., Becker, B.: Multi-objective optimisation based on the relation favour. In: Evolutionary Multi-Criterion Optimization, pp. 154–166, Switzerland, March 7–9 (2001)
3. Ishibuchi, H., Tsukamoto, N., Nojima, Y.: Evolutionary many-objective optimization: a short review. In: Proceedings of 2008 IEEE Congress on Evolutionary Computation, pp. 2424–2431, Hong Kong, June 1–6 (2008)
4. López-Jaimes, A., Coello-Coello, C.A., Chakraborty, D.: Objective reduction using a feature selection technique. In: Genetic and Evolutionary Computation, pp. 673–680, USA, July 12–16, (2008)
5. Roberts, S.A., Calamai, P.H., Hall, G.B.: Evolutionary multi-objective optimization for landscape system design. J. Geographical Syst. **13,** 299–326 (2011)
6. Roberts, S.A., Cruz-Córtes, N., Hall, B.: Evolutionary Multi-objective Optimization Design for Peri-urban Greenlands Systems: metric implementations. In: Applied Mathematics, Modeling and Computational Science, p. 195, Canada, August 26–30, (2013)
7. Saxena, D.K., Duro, J.A., Tiwari, A., Deb, K., Zhang, Q.: Objective reduction in many-objective optimization: linear and nonlinear algorithms. IEEE Trans. Evol. Comput. **17,** 77–99 (2013)
8. Singh, H.K., Isaacs, A., Ray, T.: A pareto corner search evolutionary algorithm and dimensionality reduction in many-objective optimization problems. IEEE Trans. Evol. Comput. **15,** 539–556 (2011)
9. Walker, D.J., Everson, R., Fieldsend, J.E.: Visualizing mutually nondominating solution sets in many-objective optimization. IEEE Trans. Evol. Comput. **17,** 165–184 (2013)

Evolutionary Multiobjective Optimization (EOM) Design for Peri-urban Greenlands Systems: Metric Implementations

S. A. Roberts, Nareli Cruz-Cortés and G. B. Hall

Abstract Habitat fragmentation and loss is a key issue for land-use planning and environmental policy implementation. Greenlands Systems have been proposed as one solution to this issue. This chapter discusses aspects of implementation of an evolutionary multiobjective optimization (EMO) methodology for a Greenlands System design. The application of landscape ecology principles via EMO, combined with analysis of the Pareto front of nondominated solutions and the measure of favor of these solutions provides a methodology to address the deterioration of ecological function in urban fringe areas and insights into the steps that can be taken to promote sustainable peri-urban landscapes. The results of particular landscape metrics using a real-world data set of a small study area bring into relief issues concerning the interplay between the mathematization of landscape ecology principles of design and the resulting set of estimates of the Pareto optimal solutions.

1 Introduction

Some spatial decision support problems can be framed as configuration optimization problems, a combinatorial problem. Greenlands System design is such a problem. A promising way to solve such problems is via evolutionary multiobjective optimization (EMO) methodologies.

Habitat fragmentation and loss, particularly in the urban fringe, is a key issue for land-use planning and environmental policy implementation. Greenlands Systems have been proposed as one solution to this issue [6]. The application of landscape

S. A. Roberts
Department of Geography and Environmental Studies, Wilfrid Laurier University,
Waterloo, ON, Canada
e-mail: sroberts@wlu.ca

N. Cruz-Cortés
Centro de Investigación en Computatión, Instituto Politécnico National, Mexico City, Mexico
e-mail: nareli@cic.ipn.mx

G. B. Hall
ESRI, Toronto, ON, Canada
e-mail: bhall@ESRI.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_55

ecology principles via EMO combined with a hierarchical clustering methodology are described in [12]. The EMO methodology and data set structure—primal dual multi-valued vector map (PDMVVM)—are addressed in more detail in [13]. Landscape ecologists recognize the importance to ecological function of regional scale spatial patterns of features across the landscape and they have catalogued a lexicon of spatial configurations of natural features that allow for, or promote, beneficial ecological functioning, or conversely spawn degradation of existing systems. Some of the configurations, taken from Dramstad et al. [3], that address the goal of sustainable ecological function under the threat of disruption due to urban expansion are reviewed below.

A general overview of EMO for spatial decision making is provided in Xiao [16]. Roberts et al. [13] provide theoretical and applied foundations for extending EMO into the landscape optimization domain. Other papers have assessed different aspects of the general problem domain. For example, a raster-based nondominated sorting genetic algorithm (NSGA) for forest management optimization is discussed in Ducheyne et al. [5]. Wei and Murray [14] describe EMO for facility dispersion. Finding better approximations of Pareto optimality for multiobjective solutions to routing problems are addressed in Huang et al. [7], and Huang et al. [8] introduce a related EMO methodology, artificial immune systems, for Pareto optimization of a land use allocation problem.

Path or corridor optimization problems are another area where multicriteria methods have been introduced. For example, Matisziw and Murray discuss spatial association optimization for nature reserve design [10]. EMO is used for path optimization in Mooney and Winstanley [11]. EMO is applied to the corridor location problem in Zhang and Armstrong [17]. Connectivity for natural reserve design using genetic algorithms (GAs) is described in Loonen et al. [9].

This research, and that presented in this chapter suggests that EMO-based methodologies are emerging as important tools for addressing issues of landscape configuration and decision support.

2 Configuration Design Process

Now a small spatial design problem is introduced and a sketch of how EMO techniques can be used is presented. The source data sets comprise four primal planar attributed graphs of respectively, property cadastre, soils, groundwater recharge, and ecological land classification. These inputs are combined to create the initial primal planar attributed multivalued graph and its dual. The primal graph was generated via a sequence of spatial union operations. The dual graph was created using the primal graph's pseudocentroids as vertices and the primal polygon topology to generate the adjacency edges.

Finally, preprocessing was applied to implement some design goals to create an initial data set as shown in Fig. 1, see also Table 1 for a listing of the preprocessing functions applied. The initial land-use classes have been aggregated into four classes.

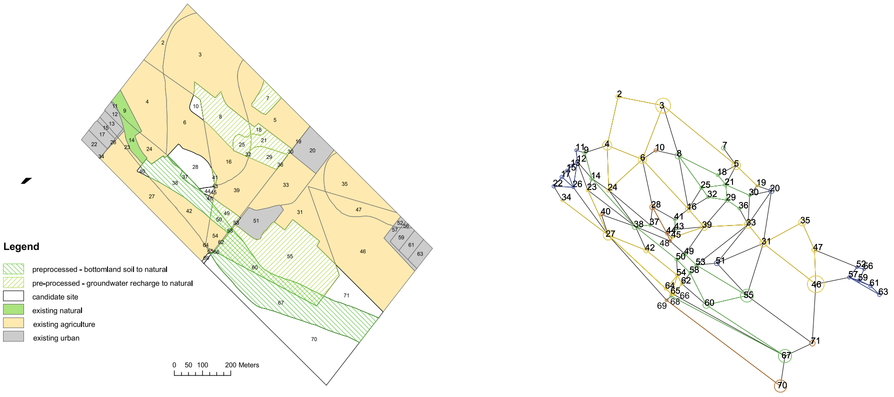


Fig. 1 The preprocessed example data set, primal planar graph (*left*) and dual graph (*right*)

Note that eight candidate sites for reassignment to create the potential design solutions were defined based on a subset of the initial land-use types that represent transitioning landscapes, mostly abandoned farmland. This makes ecological sense and considerably narrows (to $3^8 = 6561$ possible solutions) the search space for the design problem.

We coded the design problem as a many-objective optimization problem to be solved by an evolutionary multiobjective solver (see Fig. 2), in this case NSGA-II [2] which incorporates the notion of Pareto dominance to deal with multiple, potentially conflicting objectives. This methodology is enhanced by sorting the first nondominated front via a ranking based on favor ([4], chapter "Analysis and Visualization of a Many Objective Optimization Landscape Design Problem").

Formally, $<_d$, defines the relation, solution x dominates solution y and is described for a minimization problem as follows,

$$x <_d y :\Leftrightarrow (\exists i : f_i(x) < f_i(y)) \wedge (\forall j \neq i : f_j(x) \leq f_j(y)) \tag{1}$$

The set of all nondominated solutions forms the Pareto-optimal set. Further, the relation favour $<_f$ is defined as follows [4],

$$x <_f y \Leftrightarrow |\{i : f_i(x) < f_i(y)\}| > |\{i : f_i(x) < f_i(y)\}| \tag{2}$$

In our implementation we generated our ranking of the nondominated solutions by scoring each solution based on the number of other solutions it is more favoured than.

3 Evolution of Metrics and Full Enumeration Solution Results

The original alphabet of four potential partition labels: *No change, Natural, Agriculture, Urban*, was revised to three by dropping the first label because when examining the complete enumeration of solutions, solutions that contained the *No change* label

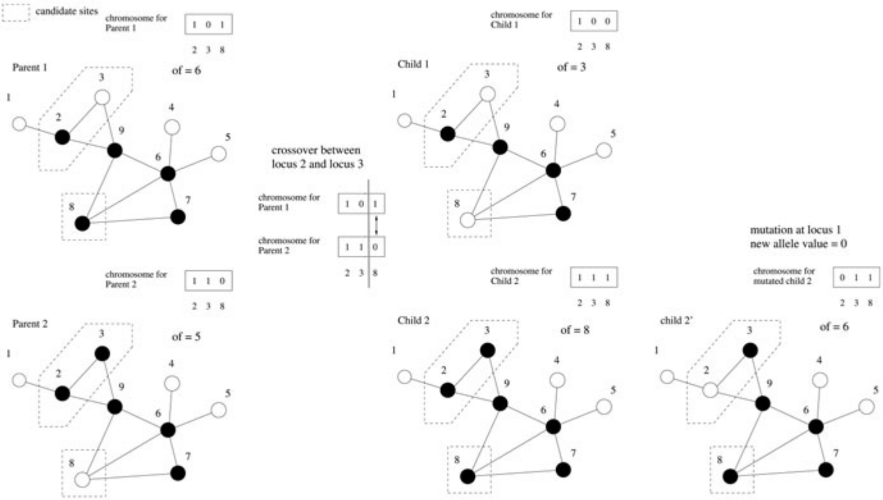


Fig. 2 A simplified example of problem coding for Genetic Algorithm solution using a binary alphabet and one objective function that measures the number of Black–Black joins

Table 1 Operationalizing Greenlands design objectives 1 (pag ≡ primal attributed graph)

Specific design objective	Constraint function or preprocessing	Name	Data	Operations on data
Include areas of natural vegetation	(Area natural)/(area total) $\geq$ min$_n$	C1	pag	Calculate area
Include areas of agri/silviculture	(Area agri/silviculture)/(area total)$\geq$ min$_a$	C2	pag	Calculate area
Include areas of urban use	(Area urban)/(area total) $\geq$ min$_u$	C3	pag	Calculate area
Vegetated riparian corridors	Preprocessed	PP1	pag	"Bottomland" soil to natural
Areas of wet soil preserved as wetlands	Preprocessed	PP2	pag	"Muck" and "Organic" soil to wetland
Vegetated ground water recharge areas	Preprocessed	PP3	pag	Candidate sites over ground water recharge areas to natural

were all Pareto dominated. Testing with an objective reducing methodology ([1], chapter "Analysis and Visualization of a Many Objective Optimization Landscape Design Problem") we found the area-based objective functions overwhelmed the other topological or structural objective functions that were key to embodying the design principles. So these objectives were recast as minimum area constraints. We also added two new objective functions related to agricultural uses (see Table 2). Further, we note that for the small study area objective **F4** could only be one value

Table 2 Operationalizing Greenlands design objectives 2 (dag ≡ dual attributed graph)

Specific design objective	Objective function to maximize	Name	Data	Operations on data
A few large patches of natural vegetation (core areas)	Mean of top 5 core area's Area Weighted Inverse Mean Shape Index	F1	p/dag	calculate area & perimeter,find subgraphs
Vegetated corridors between core areas	Mean of top 5 core area's number of vertices normalized by the total number of natural vertices	F2	p/dag	find subgraph, number of vertices
Stepping stones of natural vegetation between top 5 core areas	Number of natural vertices normalized by the total number of vertices along 10 shortest paths between top 5 core areas	F3	p/dag	find subgraph, find graph centre, shortest path, number of vertices
Natural patches in urban areas	Number of natural–urban neighbors in urban areas	F4	dag	natural-urban join count & find subgraph
Contiguous agricultural areas	Number of agricultural–agricultural neighbors	F5	dag	agric.-agric. join count
Suitable agricultural areas	Score candidate sites coded for agriculture based on soil capability	F6	dag	find subgraph, assign score
Clustered urban development	Number of urban–urban neighbors	F7	dag	urban-urban join count

and **F2** was restricted to only two values so these objectives were considered redundant by the formal analysis. However, earlier analysis suggests that this will not be the case for larger study areas.

Based on a complete enumeration (not possible with larger study areas) there are 6561 possible solutions, 47 nondominated solutions, and 9 nondominated and constrained solutions. Figure 3 displays the results of applying the optimization metrics of nondomination and favour to the complete enumerated solution space. The top three solutions using our proposed Pareto and favor ranking all display characteristics reflecting the desired design goals. Specifically, this included maintaining or expanding spatial contiguity within each of the three land-use classes and discovering and connecting "core" natural areas.

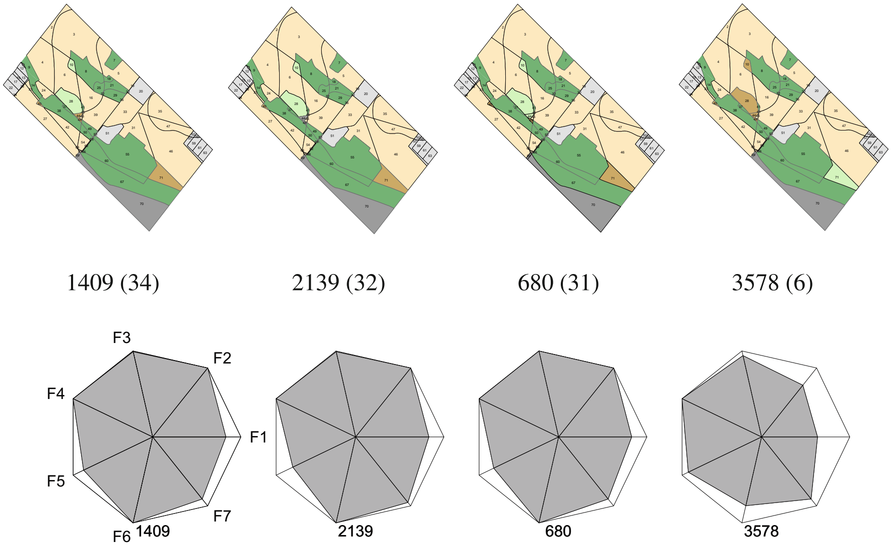


Fig. 3 Top 3 and bottom nondominated and constrained solutions, favor ranking in brackets. Note: objective function values have been normalized to the maximum values in the nondominated set to help to highlight the favor ranking results in the star plots

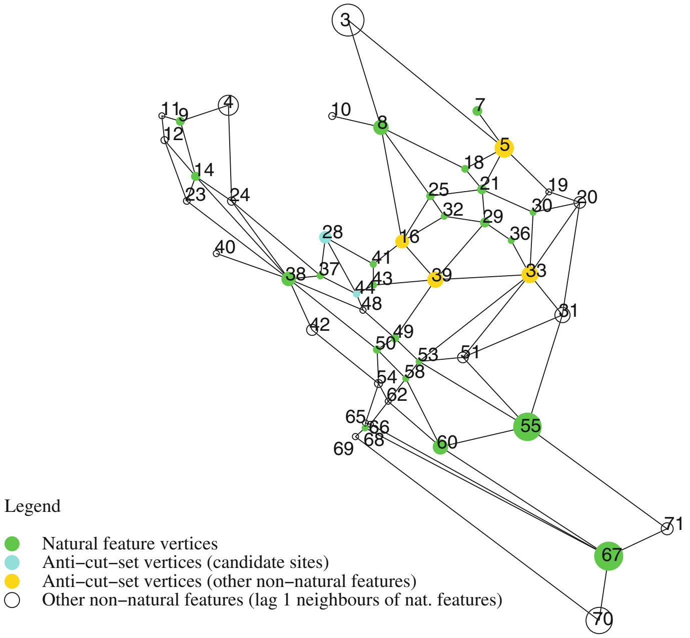

3
4
5
7
8
9
10
11
12
14
16
18
19
20
21
23
24
25
28
29
30
31
32
33
36
37
38
39
40
41
42
43
44
48
49
50
51
53
54
55
58
60
62
65
66
67
68
69
70
71
Legend
● Natural feature vertices
● Anti–cut–set vertices (candidate sites)
● Anti–cut–set vertices (other non–natural features)
○ Other non–natural features (lag 1 neighbours of nat. features)

Fig. 4 Anti-cut-set vertices as a basis for “directed mutation”

4 Conclusions

We have described in this chapter the results of applying recently developed tools for objective function reduction analysis and favor ranking to enhance both problem definition and solution techniques for EMO approaches to the configuration optimization problem of Greenlands System design. We will be testing our revised objectives, again using NSGA-II, to compare the GA's performance in finding the true Pareto front that is explicitly available with our small data set but will not be for larger problem domains. In progress is implementation of anti-cut-set vertex based "directed mutation" (see Fig. 4). This variation is meant to more explicitly utilize the topological information of the coded solutions than the current approach. Next, is revising code to parallelize the implementation of NSGA-II and include favor ranking to run on general purpose graphics processor unit (GPGPU) based computer and testing with larger data sets.

References

1. Brockoff, D., Zitzler, E.: Objective reduction in evolutionary multiobjective optimization: theory and applications. Evol. Comput. **17**(2), 135–166 (2009)
2. Deb, K., Pratap, A., Agarak, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **8**(2), 182–197 (2002)
3. Dramstad, W.E., Olson, J.D., Forman, R.T.T.: Landscape Ecology Principles in Landscape Architecture and Land-Use Planning. Harvard Graduate School of Design: Island Press, American Society of Landscape Architects, Washington, DC, (1996)
4. Drechsler, N., Drechsler, R., Becker, B.: Multi-objective Optimisation Based on Relation Favour, Evolutionary Multi-Criterion Optimisation, LNCS, pp. 154–166, Springer (2001)
5. Ducheyne, E.I., De Wulf, R.R., De Baets, B.: A spatial approach to forest-management optimization: linking GIS and multiple objective genetic algorithms. Int. J. Geogr. Inf. Sci. **20**(8), 917–928 (2006)
6. Forman, R.T.T.: Land Mosaics: The Ecology of Landscapes and Regions. Cambridge University Press Cambridge, U.K. (1997)
7. Huang, B., Fery, P., Xue, L., Wang, Y.: Seeking the Pareto front for multiobjective spatial optimization problems. Int. J. Geogr. Inf. Sci. **22**(**5**), 507–526 (2008)
8. Huang, K., Lui, X., Li, X., Liang, J., He, S.: An improved artificial immune system for seeking the Pareto front of land-use allocation problem in large areas. Int. J. Geogr. Inf. Sci. **27**(5), 922–946 (2013)
9. Loonen, W., Heuberger, P.S.C., Bakema, A.H., Schot, P.: Improving the spatial coherence of nature using genetic algorithms. Environ. Plan. B **34,** 369–378 (2007)
10. Matisziw, T.C., Murray, A.T.: Promoting species persistence through spatial association optimization in nature reserve design. J. Geogr. Syst. **8,** 289–305 (2006)
11. Mooney, P., Winstanley, A.: An evolutionary algorithm for multicriteria path optimization problems. Int. J. Geogr. Inf. Sci. **20**(4), 401–423 (2006)
12. Moulton, C.M., Roberts, S.A., Calamai, P.H.: Hierarchical clustering of multiobjective optimization results to inform land use decision making. Urban Reg. Inf. Syst. J. **21**(2), 25–37 (2009)
13. Roberts, S.A., Calamai, P.H., Hall, G.B.: Evolutionary multiobjective optimization results for landscape system design. J. Geogr. Syst. **13,** 299–326 (2011)

14. Wei, R., Murray, A.T.: A multi-objective evolutionary algorithm for facility dispersion under conditions of spatial uncertainty. J. Oper. Res. Soc. (published online 22 May 2013) (2013)
15. Xiao, N., Bennett, D.A., Armstrong, M.P.: Interactive evolutionary approaches to multiobjective spatial decision making: a synthetic review. Comput. Environ. Urban Syst. **31,** 232–252 (2007)
16. Xiao, N.: A unified conceptual framework for geographical optimization using evolutionary algorithms. Ann. Assoc. Am. Geogr. **98**(4), 795–817 (2008)
17. Zhang, X., Armstrong, M.P.: Genetic algorithms and the corridor location problem: multiple objectives and alternate solutions. Environ. Plan. B **35,** 148–168 (2008)

Effect of Boundary Absorption on Dispersion of a Solute in Pulsatile Casson Fluid Flow

B. T. Sebastian and P. Nagarani

Abstract The generalized dispersion model is used to study the dispersion process in unsteady flow in a tube with wall absorption by modeling the flowing fluid as Casson fluid. According to this model, the entire dispersion process is expressed in terms of three transport coefficients viz., the absorption, convection, and dispersion coefficients. This study brings out the effects of pulsatility, yield stress and wall absorption on these three transport coefficients. It is observed that the convection and the dispersion coefficients are dependent on absorption parameter, yield stress, pressure fluctuating component, and frequency parameter whereas the absorption coefficient depends only on wall absorption parameter. This study can be used to understand dispersion process in blood flows.

1 Introduction

The longitudinal dispersion of a tracer in a tube has many applications in the fields of chemical engineering, environmental dynamics, and biomedical engineering. Taylor [6] was first to initiate the study on contaminant dispersion in a circular tube flow and showed that when a soluble substance is introduced into a fluid moving slowly and steadily through a circular tube it spreads out due to the combined action of molecular diffusion and the variation of velocity over the cross section. Aris [1] extended this by the method of moments including the effect of axial molecular diffusion. These theories are applicable only for large time after the introduction of solute and did not provide any idea about variation of the dispersion coefficient immediately after the injection of solute. Gill and Sankarasubramanian [3] developed a method to study

B. T. Sebastian (✉) · P. Nagarani
The University of the West Indies, Mona Campus, Kingston, Jamaica, W.I.
e-mail: binil5@yahoo.com

B. T. Sebastian
University of Technology, Kingston, Jamaica

P. Nagarani
e-mail: nagarani_ponakala@yahoo.co.in

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_56

the dispersion of a solute in a tube and this model is widely called as a generalized dispersion model, which holds for all times after the solute injection. Later this model is extended in the case of wall absorption by Sankarasubramanian and Gill [4]. They showed that the three effective transport coefficients namely absorption, convection, and dispersion coefficient are affected by interphase mass transfer. Dash et al. [2] gave a model to understand the dispersion process in a Casson fluid by considering the flowing fluid as steady and showed that the dispersion coefficient in the case of Cason fluid depends not only on time but also on yield stress. They also discussed the applications of their study in understanding the dispersion process in blood flows.

The existed models in the literature explain the effects of non-Newtonian rheology on dispersion of solute but not the other properties of blood flow. Blood flow in arteries and veins exhibits not only the non-Newtonian nature but also many other fluid dynamic complexities such as pulsatility, curvature, branching, and elasticity of the walls. The dispersion of any solute in blood flow is affected by these phenomena as well as the wall reaction mechanisms and the multiphase character of the blood. Hence, in this chapter, an attempt is made to study the dispersion process in a tube with wall absorption by considering the flow as unsteady and flowing fluid as Casson fluid. The purpose of this study is to explore the combined effects of yield stress, Womersley parameter, fluctuating pressure component, and absorption parameter on dispersion coefficient in a Casson fluid flowing through a tube.

2 Mathematical Formulation

we considered axisymmetric, fully developed, pulsatile flow in a pipe of radius "a" by modeling the flow as a Casson fluid flow. We assumed that the rate of disappearance of solute at the tube wall is due to an irreversible first-order reaction catalyzed by the wall and is proportional to the solute concentration of the wall. The unsteady convective diffusion equation that describes the local concentration C of a solute as a function of axial distance z, radial distance r, and time t in the nondimensional form can be written as follows:

$$\frac{\partial C}{\partial t} + w(r,t)\frac{\partial C}{\partial z} = \left(\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial}{\partial r}) + \frac{1}{Pe^2}\frac{\partial^2}{\partial z^2}\right)C \tag{1}$$

with the nondimensional variables as follows:

$$C = \frac{\overline{C}}{C_0}, w = \frac{\overline{w}}{w_0}, r = \frac{\overline{r}}{a}, z = \frac{D_m\overline{z}}{a^2 w_0}, t = \frac{D_m\overline{t}}{a^2}, Pe = \frac{a w_0}{D_m} \tag{2}$$

where w is the nondimensional axial velocity of the fluid, D_m is the coefficient of molecular diffusion (molecular diffusivity) which is assumed to be constant, C_0 is the reference concentration, w_0 is the characteristic velocity and Pe is the Peclet number. The variables with bar indicate the corresponding variables in dimensional form. For the slug input of solute length z_s under consideration, the initial and boundary conditions in dimensionless form for the given model will be of the form:

$$C(0,z,r) = \begin{cases} 1 & \text{if } |z| \leq \frac{z_s}{2} \\ 0 & |z| > \frac{z_s}{2}, \end{cases} \tag{3}$$

$$\frac{\partial C}{\partial r}(t,z,0) = 0, \tag{4}$$

$$\frac{\partial C}{\partial r}(t,z,1) = -\beta C, \tag{5}$$

$$C(t,\infty,r) = 0, \tag{6}$$

where β is the wall absorption parameter.

The constitutive equation for a Casson fluid relating the stress (τ) and shear rate $\left(\frac{\partial w}{\partial r}\right)$ in nondimensional form is given by

$$\tau^{\frac{1}{2}} = \tau_y^{\frac{1}{2}} + \left(-\frac{\partial w}{\partial r}\right)^{\frac{1}{2}} \quad \text{if } \tau \geq \tau_y, \tag{7}$$

$$\frac{\partial w}{\partial r} = 0 \, \text{if } \tau \leq \tau_y, \tag{8}$$

where $\tau_y = \dfrac{\overline{\tau}_y}{\mu(w_0/a)}$ and $\tau = \dfrac{\overline{\tau}}{\mu(w_0/a)}$ are the nondimensional yield stress and shear stress, respectively. The above relations correspond to vanishing of velocity gradient in the region where the shear stress is less than the yield stress which implies a plug flow for $\tau \leq \tau_y$. The nondimensional velocity distribution for axisymmetric, fully developed, unsteady flow of a Casson fluid in tube is given by [5] as follows:

$$\begin{aligned} w = w_{-} = w_p = \frac{1}{2}p(t)\left\{1 - \frac{8}{3}r_p^{\frac{1}{2}} + 2r_p - \frac{1}{3}r_p^2\right\} \\ -\alpha^2\frac{p'(t)}{32}\left\{3 - \frac{1144}{147}r_p^{\frac{1}{2}} + \frac{320}{63}r_p + \frac{4}{3}r_p^2 - \frac{16}{9}r_p^{\frac{5}{2}} + \frac{65}{441}r_p^4\right\} \\ \text{if } 0 \leq r \leq r_p, \end{aligned} \tag{9}$$

$$w = w_{+} = \frac{1}{2}p(t)\left\{\left(1-r^2\right) - \frac{8}{3}r_p^{\frac{1}{2}}\left(1-r^{\frac{3}{2}}\right) + 2r_p\left(1-r\right)\right\}$$
$$-\frac{\alpha^2}{2}p^{'}(t)\left\{\frac{3}{16} - \frac{r^2}{16}\left(4-r^2\right) - \frac{r_p^{\frac{1}{2}}}{16}\left[\frac{1144}{147} - \frac{16}{3}\left(r^2 + r^{\frac{3}{2}}\right) + \frac{424}{147}r^{\frac{7}{2}}\right]\right.$$
$$\left. + \frac{r_p}{16}\left[\frac{320}{63} + \frac{128}{63}r^3 - \frac{64}{9}r^{\frac{3}{2}}\right]\right\} \qquad if\ r_p \le r \le 1, \tag{10}$$

where $r_p = \dfrac{\tau_y}{p(t)}$ is the dimensionless plug radius and $p(t) = 1 + e\cos\alpha^2 Sct$. Also the subscripts "$-$" and "$+$" corresponds the values for plug flow and shear flow, respectively and $\alpha = \sqrt{\frac{\omega a^2}{\nu}}$ represents the Womersley parameter, $Sc = \dfrac{\nu}{D_m}$ represents the Schmidt number, e is the amplitude of the pressure fluctuating component.

The solution of the convective diffusion Eq. (1) along with the given set of initial and boundary conditions (3–6) by following the analysis of [3] can be assumed as follows:

$$\sum_{i=0}^{\infty} f_i(t,r)\frac{\partial^i C_m}{\partial z^i}, \tag{11}$$

where the dimensionless mean concentration C_m is defined as follows:

$$C_m = 2\int_0^1 Cr\,dr. \tag{12}$$

Multiplying Eq. (1) by $2r$ and integrating with respect to r from 0 to 1, we get

$$\frac{\partial C_m}{\partial t} = \sum_{i=0}^{\infty} K_i(t)\frac{\partial^i C_m}{\partial z^i} \tag{13}$$

with transport coefficients K_i's as function of time t and

$$K_i(t) = \frac{\partial_{i2}}{Pe^2} - 2\int_0^1 f_{i-1}(t,r)w(t,r)r\,dr + 2\frac{\partial f_i}{\partial r}(t,1),\ i = 0,1,2,3\ldots, \tag{14}$$

where δ_{ij} denotes Kronecker delta and $K_0(t)$, $K_1(t)$, and $K_2(t)$ are called as the absorption coefficient, convection coefficient, and dispersion coefficient, respectively. Also the following set of differential equations for f_n is obtained as follows:

$$\frac{\partial f_n}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f_n}{\partial r}\right) - w(t,r)f_{n-1} + \frac{1}{Pe^2}f_{n-2} - \sum_{i=0}^{n} K_i f_{n-i} \quad n = 0,1,2,\ldots. \tag{15}$$

The initial and boundary conditions are obtained from Eqs. (3–6) as follows:

$$f_n(0,r) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{for } n = 1, 2, 3 \ldots, \end{cases} \tag{16}$$

$$\frac{\partial f_n}{\partial r}(t,0) = 0, \tag{17}$$

$$\frac{\partial f_n}{\partial r}(t,1) = -\beta f_n(t,1). \tag{18}$$

In order to solve the transport coefficient one has to solve $f_n s$ simultaneously. These coupled equations are not conformable to an analytic solution, so a finite difference scheme is used to study the dispersion phenomena and is explained in Sect. 3. By neglecting terms involved K_3, K_4, etc., in Eq. (13) and solving we can get the expression for C_m.

3 Numerical Scheme

Equation (15) for $n = 0, 1, 2$ for f_n's are discretized in radial direction r and time t. The Crank–Nicolson method is applied for each time step. The finite difference scheme for derivatives and other terms are written at the mesh (i, j), where $0 \leq j \leq m$ and $0 \leq i \leq n$. The resultant finite difference equations become linear simultaneous equations with a tridiagonal matrix in the form $A_i f_n(i+1, j+1) + B_i f_n(i, j+1) + C_i f_n(i-1, j+1) = D_i$, where A_i, B_i, C_i, and D_i are the matrix elements. This tridiagonal matrices can be solved by using the Gauss Seidel method with the help of initial and boundary conditions.

4 Results and Discussion

The effect of yield stress, Womersley parameter, fluctuating pressure component, and absorption parameter on dispersion coefficient is analyzed. From Fig. 1a–d, it can be seen that due to the oscillatory flow the dispersion coefficient changes cyclically and initially increase with time. From Fig. 1b, c one can observe that fluctuations and the magnitude of K_2 increase with e and also as β increases the dispersion coefficient K_2 decreases. We also observed that as the yield stress increases the amplitude of the fluctuations of K_2 decreases.

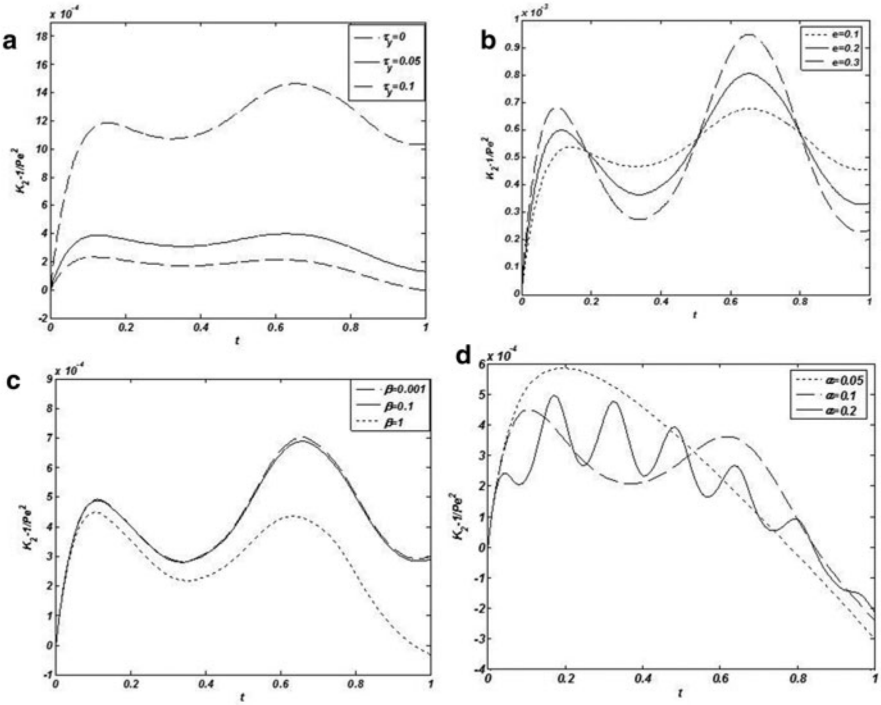


Fig. 1 Variation of dispersion coefficient $K_2$with t when $Pe = Sc = 1000$ for different **a** τ_y for $e = 0.1$, $\beta = 1$, and $\alpha = 0.1$ **b** e for $\tau_y = 0.02$, $\beta = 1$, and $\alpha = 0.1$ **c** β for $\tau_y = 0.05$, $e = 0.2$, and $\alpha = 0.1$ **d** α for $\tau_y = 0.05$, $e = 0.2$, and $\beta = 1$

5 Conclusions

The expression for dispersion coefficient is obtained for dispersion of a solute in Casson fluid flow with wall absorption by using the generalized dispersion model. The dispersion coefficient has been found to depend on yield stress, absorption parameter, frequency parameter, and the fluctuating component.

References

1. Aris, R.: On the dispersion of solute in a fluid through a tube. Proc. R. Soc. Lond. A **235,** 67–77 (1956)
2. Dash, R.K., Jayaraman, G., Mehta, K.N.: Shear augmented dispersion of a solute in a Casson fluid flowing in a conduit. Ann. Biomed. Eng. **28,** 373–385 (2000)
3. Gill, W.N., Sankarasubramanian, R.: Exact analysis of unsteady convective diffusion. Proc. R. Soc. Lond. A **316,** 341–350 (1970)

4. Rohlf, K., Tenti, G.: The role of the Womersley number in pulsatile blood flow: a theoretical study of the Casson model. J. Biomech. **34,** 141–148 (2001)
5. Sankarasubramanian, R., Gill, W.N. : Unsteady convective diffusion with interphase mass transfer. Proc. R. Soc. Lond. A **333,** 115–132 (1973)
6. Taylor, G.I.: Dispersion of soluble matter in solvent flowing slowly through a tube. Proc. R. Soc. Lond. A **219,** 186–203 (1953)

Stability Analysis of a Human–*Phlebotomus papatasi*–Rodent Epidemic Model

Schehrazad Selmane

Abstract Cutaneous leishmaniasis (CL) represents a serious public health problem in Algeria. In the aim to understand the transmission dynamics of CL in the human–*Phlebotomus papatasi*–rodent cycle, and to improve the preventive strategies set up in Algeria, we developed a deterministic model for the transmission dynamics of the disease. The model includes an incidental host for human which acts only as a sink of infection, a primary reservoir host for rodent which acts as a source and a sink of infection, and a secondary reservoir host for *P. papatasi* which have a role in transmission by acting as the liaison between incidental host and primary reservoir. The global stability of the equilibria of the proposed model shows that the threshold conditions for disease persistence are completely determined by the reproduction number and do not explicitly include parameters relating to the dynamic transmission in the incidental hosts, which means that the disease becomes endemic if it persists endemically in the primary reservoir hosts, and therefore the control measures should be directed towards reservoir hosts. This is illustrated via numerical simulations of the model using parameters generated from data from M'Sila province in Algeria.

1 Introduction

Leishmaniasis is a disease caused by protozoan parasites that belong to the genus Leishmania and is transmitted by the bite of certain species of sandflies. The only proven vectors of human disease are female sandflies (Phlebotomus species in the Old World, Lutzomyia species in the New World). They acquire Leishmania parasites and infection starts when they feed on an infected mammalian host in search of a blood meal. The leishmaniases can be grouped into two broad categories according to the source of human infection: zoonotic leishmaniases, in which the reservoir hosts are wild animals, commensals or domestic animals, and anthroponotic leishmaniases, in which the reservoir host is human. There are four main types of the

S. Selmane (✉)
LIFORCE, Faculty of Mathematics, The University of Science and
Technology Houari Boumediene, Algiers, Algeria
e-mail: cselmane@usthb.dz

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_57

disease: cutaneous leishmaniasis (CL), the most common form of leishmaniasis; diffuse cutaneous leishmaniasis (DCL), a chronic form of leishmaniasis and difficult to treat; mucocutaneous leishmaniasis (MCL), the most feared form of CL because it produces destructive and disfiguring lesions of the face; and visceral leishmaniasis (VL), also known as kala-azar, if left untreated can have a fatality rate as high as 100 % within 2 years [4].

Leishmaniasis is endemic in 98 countries, spread over four continents, with more than 350 million people at risk. An estimated incidence of 2 million new cases per year (0.5 million of VL and 1.5 million of CL). VL causes an estimated over 50,000 deaths annually, a rate surpassed among parasitic diseases only by malaria. More than 90 % of the burden of VL is concentrated in Bangladesh, Brazil, Ethiopia, India, Nepal, and Sudan. Up to 90 % of cases of CL occur in Afghanistan, Algeria, the Islamic Republic of Iran, Saudi Arabia and the Syrian Arab Republic and in Bolivia, Brazil, Colombia, Nicaragua, and Peru. Climatic, socioeconomic and other environmental changes could expand the geographical range of the vectors and leishmaniasis transmission in the future [4].

CL represents a serious public health problem in Algeria; three CL outbreaks occurred between 2004 and 2006, with, respectively, 14,822, 25,511, and 14,714 cases [5]. In 1984–1985, only few cases were reported, probably because of the application of the insecticide dichlorodiphenyltrichloroethane (DDT) in 1983 against malaria, but since 1986, the number of cases has risen to over more than 2000 cases yearly. The disease used to be mainly endemic in the sub-Saharan steppe, however, a geographical spread towards the north and west has taken place recently. The human infection is caused *Phlebotomus papatasi*, *Phlebotomus perniciosus*, *Phlebotomus sergenti*, and *Phlebotomus perfiliewi* and the disease occurs in two clinical forms: VL and CL. The notification of leishmaniasis became mandatory in 1979, and is under surveillance since 1985. A national leishmaniasis control program for VL and CL has been set in place in 2006 and care for leishmaniasis are provided for free in high incidence regions [1].

In the aim to understand the transmission dynamics of CL in the human–*P. papatasi*–rodent cycle, and to improve the preventive strategies set up in Algeria, we present a deterministic model. The model includes an incidental host for human, a primary reservoir host for rodent, and a secondary reservoir host for *P. papatasi*. The global stability of the equilibria of the proposed model is carried out in Sect. 2. Numerical simulations of the model using parameters generated from data from M'Sila province in Algeria and conclusion are reported in Sect. 3.

2 Model : Interactions in CL

The model includes:

- an incidental host for human which acts only as a sink of infection, that is, human cannot transmit the disease

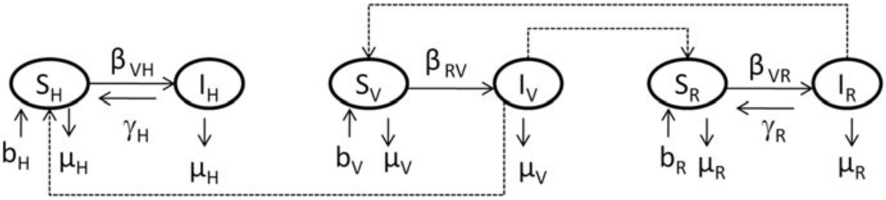


Fig. 1 Flows between the compartments of the model

- a primary reservoir host for rodent which acts as a source and a sink of infection, that is, rodent can transmit the pathogen to new hosts
- a secondary reservoir host for *P. papatasi*, which has a role in transmission by acting as liaison between incidental host and primary reservoir.

The dynamics of transmission is bidirectional between vectors and reservoir hosts, that is, they are all both source and sink of infection. The model is schematically illustrated in Fig. 1; the interactions of the six compartments are specified by the normalized system of nonlinear differential Eq. (1), and parameters are described in Table 1.

$$\begin{cases} \dot{S_H} = b_H - \beta_{VH} m_H I_V S_H + \gamma_H I_H - \mu_H S_H \\ \dot{I_H} = \beta_{VH} m_H I_V S_H - (\gamma_H + \mu_H) I_H \\ \dot{S_V} = b_V - \beta_{RV} S_V I_R - \mu_V S_V \\ \dot{I_V} = \beta_{RV} S_V I_R - \mu_V I_V \\ \dot{S_R} = b_R - \beta_{VR} m_R I_V S_R + \gamma_R I_R - \mu_R S_R \\ \dot{I_R} = \beta_{VR} m_R I_V S_R - (\gamma_R + \mu_R) I_R, \end{cases} \tag{1}$$

where $m_H = \frac{N_V}{N_H}$ and $m_R = \frac{N_V}{N_R}$.

System (1) has one disease-free equilibrium (DFE): $E_0^* = \left(\frac{b_V}{\mu_V}, 0, \frac{b_H}{\mu_H}, 0, \frac{b_R}{\mu_R}, 0\right)$. The basic reproduction number is computed using the next generation operator [2]

$$\mathcal{R}_0 = \sqrt{\left(\frac{b_R}{\mu_R} \frac{m_R \beta_{RV}}{(\gamma_R + \mu_R)}\right) \left(\frac{b_V}{\mu_V} \frac{\beta_{VR}}{\mu_V}\right)}.$$

By analyzing the linear part of system (1), the local stability properties are established. The stability of the DFE is achieved through the determination of the sign of the eigenvalues of the jacobian matrix of system (1) evaluated at DFE; four negative eigenvalues are straightforwardly determined $-\mu_H$, $-\gamma_H - \mu_H$, $-\mu_V$, $-\mu_R$ and the two remaining satisfy:

$$\lambda^2 + (\mu_V + \gamma_R + \mu_R) \lambda + \left(1 - \mathcal{R}_0^2\right) \mu_V (\gamma_R + \mu_R) = 0.$$

As $\mu_V + \gamma_R + \mu_R > 0$, thus according to the Routh–Hurwitz criterion, if $\mathcal{R}_0 < 1$ the roots are with negative real parts and thus the DFE is locally asymptotically stable. Otherwise, the DFE is unstable and an epidemic is triggered. Moreover, the conditions $(H1)$ and $(H2)$ [2] are met, hence, the DFE is globally asymptotic stable whenever $\mathcal{R}_0 < 1$.

Setting the derivatives equal to zero in system (1) and solving the corresponding algebraic system, one gets a unique endemic equilibrium $E^* = (S_H^*, I_H^*, S_V^*, I_V^*, S_R^*, I_R^*)$

$$\begin{array}{lll} S_V^* = \frac{b_V}{\mu_V} - I_V^* & S_H^* = \frac{b_H}{\mu_H}\frac{1}{1+k_1 I_V^*} & S_R^* = \frac{b_R}{\mu_R}\frac{1}{1+k_2 I_V^*} \\ I_V^* = \frac{\mu_R \mu_V}{k_2(b_R \beta_{RV} + \mu_R \mu_V)}\left(\mathcal{R}_0^2 - 1\right) & I_H^* = \frac{b_H}{\mu_H}\frac{k_1 I_V^*}{1+k_1 I_V^*} & I_R^* = \frac{b_R}{\mu_R}\frac{k_2 I_V^*}{\left(1+k_2 I_V^*\right)} \end{array}$$

where $k_1 = \frac{\beta_{VH} m_H}{\gamma_H + \mu_H}$ and $k_2 = \frac{\beta_{VR} m_R}{\gamma_R + \mu_R}$ if $\mathcal{R}_0 > 1$, and no equilibria with positive coordinates if $\mathcal{R}_0 \leq 1$. All the eigenvalues of the Jacobian at the endemic equilibrium E^* are negative real whenever $\mathcal{R}_0 > 1$, which ensure the local asymptotic stability of E^*. Indeed, four negative eigenvalues are straightforwardly determined $-\mu_H$, $-\mu_V$, $-\mu_R$, $-\gamma_H - \mu_H - \beta_{VH} m_H I_V^*$ and the two remaining eigenvalues satisfy:

$$\lambda^2 + \left[(\gamma_R + \mu_R)\left(1 + k_2 I_V^*\right) + \frac{\beta_{RV} b_R k_2 I_V^*}{\mu_R \left(1 + k_2 I_V^*\right)} + \mu_V\right]\lambda + (\gamma_R + \mu_R)\,\mu_V \left(\mathcal{R}_0^2 - 1\right) = 0.$$

As the coefficients are positive for $\mathcal{R}_0 > 1$, thus according to the Routh–Hurwitz criterion, the roots are with negative real parts.
The global stability of the endemic equilibrium E^* was established using the following Lyapunov function $V := V(S_H, I_H, S_V, I_V, S_R, I_R)$

$$V = \left(S_H - S_H^* - S_H^* \log\frac{S_H}{S_H^*}\right) + \left(I_H - I_H^* - I_H^* \log\frac{I_H}{I_H^*}\right) + \left(S_V - S_V^* - S_V^* \log\frac{S_V}{S_V^*}\right) + \left(I_V - I_V^* - I_V^* \log\frac{I_V}{I_V^*}\right) + \left(S_R - S_R^* - S_R^* \log\frac{S_R}{S_R^*}\right) + \left(I_R - I_R^* - I_R^* \log\frac{I_R}{I_R^*}\right).$$

Using system (1) and substituting $S_H = S_H - S_H^*$, $S_V = S_V - S_V^*$, $S_R = S_R - S_R^*$, $I_H = I_H - I_H^*$, $I_V = I_V - I_V^*$, $I_R = I_R - I_R^*$, the Lyapunov derivative takes this form: $\frac{dV}{dt} = A + B$ where

$$\begin{aligned} A = &-\frac{\left(S_H - S_H^*\right)^2}{S_H}\left(\beta_{VH} m_H \left(I_V - I_V^*\right) + \mu_H\right) - (\gamma_H + \mu_H)\frac{\left(I_H - I_H^*\right)^2}{I_H} \\ &-\frac{\left(S_V - S_V^*\right)^2}{S_V}\left(\beta_{RV}\left(I_R - I_R^*\right) + \mu_V\right) - \mu_V \frac{\left(I_V - I_V^*\right)^2}{I_V} \\ &-\frac{\left(S_R - S_R^*\right)^2}{S_R}\left(\beta_{VR} m_R \left(I_V - I_V^*\right) + \mu_R\right) - (\gamma_R + \mu_R)\frac{\left(I_R - I_R^*\right)^2}{I_R} \end{aligned}$$

Table 1 Values of parameters used in the simulations

Parameter	Sandflies	Humans	Rodents
Biting rate	$\beta_{RV} = 1/14$	$\beta_{VH} = 0.3/14$	$\beta_{VR} = 1/14$
Recovery rate	–	$\gamma_H = 1/6$ per week	$\gamma_R = 1/4$ per year
Death rate	$\mu_V = 0.42$ per day	$\mu_H = 1/70$ per year	$\mu_R = 1/5$ per year
Total population	$N_V = 5000$	$N_H = 918557$	$N_R = 500$
Birth rate	$b_V = a_0 \sin \frac{2\pi}{365}(t - a_1) + a_2$	$b_H = \mu_H$	$b_R = \mu_R$

$$\begin{aligned} B =& \frac{S_H - S_H^*}{S_H}\left[b_H + \gamma_H\left(I_H - I_H^*\right)\right] + \frac{I_H - I_H^*}{I_H}\beta_{VH}m_H\left(I_V - I_V^*\right)\left(S_H - S_H^*\right) \\ &+ b_V\frac{S_V - S_V^*}{S_V} + \frac{I_V - I_V^*}{I_V}\beta_{RV}\left(S_V - S_V^*\right)\left(I_R - I_R^*\right) \\ &+ \frac{S_R - S_R^*}{S_R}\left[b_R + \gamma_R\left(I_R - I_R^*\right)\right] + \frac{I_R - I_R^*}{I_R}\beta_{VR}m_R\left(I_V - I_V^*\right)\left(S_R - S_R^*\right). \end{aligned}$$

It follows that $\frac{dV}{dt} \leq 0$ if and only if $B \leq 0$; thus, the largest compact invariant set in $\{(S_H, I_H, S_V, I_V, S_R, I_R) \in \Omega : \frac{dV}{dt} = 0\}$ is a singleton $\{E^*\}$ and hence E^* is globally asymptotically stable in the region $\Omega = \{(S_H, I_H, S_V, I_V, S_R, I_R) \in \mathbb{R}_+^6 : S_H + I_H = S_V + I_V = S_R + I_R = 1\}$.

3 Numerical Simulation and Conclusion

The global stability of the equilibria of the proposed model shows that the threshold conditions for disease persistence are completely determined by the basic reproduction number $\mathcal{R}_0$. The latter do not explicitly include parameters relating to the dynamic transmission in human population, which means that the disease becomes endemic if it persists endemically in the primary reservoir hosts. This is to be expected, because the introduction of infected incidental hosts will not cause infections in susceptible vectors, and therefore, will not produce new infections neither in susceptible incidental hosts nor in primary hosts. Consequently, the control measures should be rather directed towards reservoir hosts, namely, sandflies and rodents. Another consequence of results is the inadequacy of the definition of $\mathcal{R}_0$ as the number of secondary cases produced by the introduction of a primary case in multispecific systems where not all populations play the role of both source and sink of parasites. That is, in host populations that only act as sinks of parasites, the generation of new infections necessarily depends on the hosts that act as reservoirs of the parasites.

The obtained results are illustrated via numerical simulations of the model using parameters presented in Table 1; some of which were generated from data from M'Sila province, and others were taken from related works. We ran the simulations with the initial conditions: $I_H = 0$, $I_V = 0$, and $I_R = 50/N_R$. A bifurcation diagram describing bifurcation at $\mathcal{R}_0 = 1$ is depicted in Fig. 2, and the general behavior of the model is shown in Fig. 3.

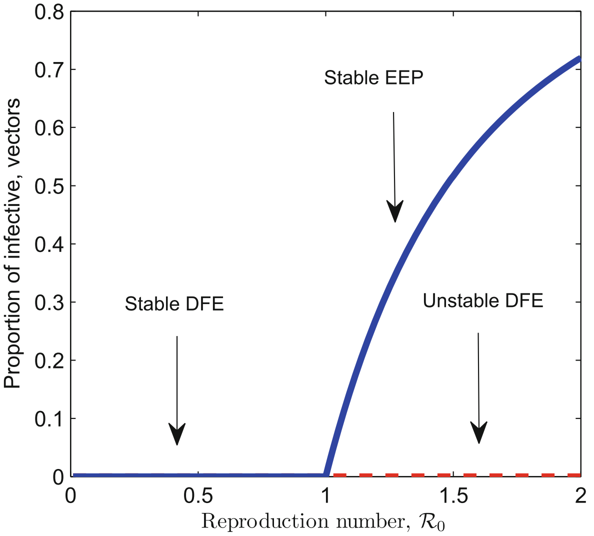


Fig. 2 The diagram of forward bifurcation

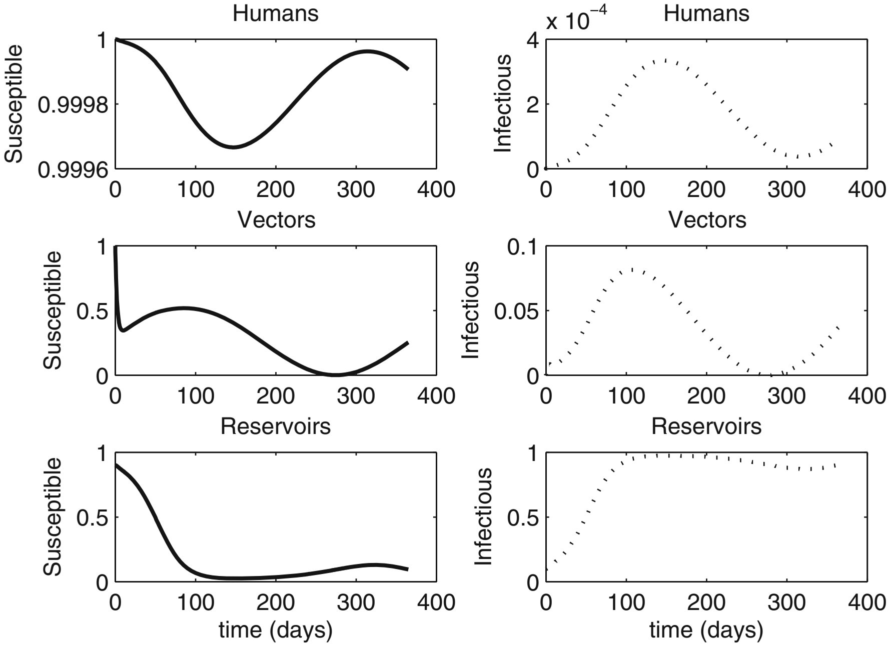


Fig. 3 The proportion of susceptible and infected hosts is plotted as function of time for $a_0 = 1/8$, $a_1 = 90 - (365/4)$, $a_2 = 1/8$

References

1. Bachi, F.: Aspects épidémiologiques et cliniques des leishmanioses en Algérie. La Lettre de l'infectiologue **21**(1), 9–15 (2006)
2. Castillo-Chavez, C., Feng, Z., Huang, W.: On the computation Ro and its role on global stability. In: Castillo-Chavez, C., Blower, S., van den Driesschie, P., Kirschner, D., Yakubu, A.A. (eds) Mathematical approaches for emerging and reemerging infectious diseases: an introduction (IMA Volumes in Mathematics and its Applications, vol 125). Springer-Verlag, Berlin-Heidelberg-New York, pp 229–250 (2002)
3. Chaves, L.F., Hernandez, M.J., Dobson, A.P., Pascual, M.: Sources and sinks: revisiting the criteria for identifying reservoirs for American cutaneous leishmaniasis. Trends Parasitol **23**, 311 (2007)
4. Control of the leishmaniasis report of a meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva, 22–26 March. (WHO technical report series ; no. 949) (2010)
5. Anonymes (2000–2010) Institut national de la santé publique (INSP). REM (relevé épidémiologique mensuel). http://www.ands.dz/insp/remhtml

Computational Thinking and Simulation in Teaching Science and Mathematics

Hasan Shodiev

Abstract Characteristics of scientific phenomenon are commonly investigated using mathematical tools in science and engineering to develop our conceptual understanding. However, computational thinking (CT) and modeling with simulations can result in a more advanced understanding of scientific concepts and offer an effective learning experience for students with various backgrounds. In this chapter, we show how a simulation tool, Scratch, can be used to unfold the abstract side of science through project-based visualizations in fun and engaging ways. It can be an effective approach in attracting young talented students to science and technology by motivating their natural imagination to probe scientific abstraction.

1 Introduction

The educational system is lacking in progress in implementing the computational approach to understand nature and technology. However, the science community has developed a novel method of solving problems by simulating various phenomena in many science and engineering disciplines. This offers us new answers to scientific questions that are different from theory and experimentation. Computational thinking (CT) in addition to critical thinking is very important when utilizing the computational approach. This can be illustrated in Fig. 1, where the diagram shows a CT path parallel to critical thinking as an integral part of obtaining solutions.

CT emerged as a new paradigm alongside mathematical, physical, musical, and other types of thinking after the availability of computers. CT in problem solving was first introduced by Dr. Seymour Papert [4]. In 1971, Dr. Papert showed the use of CT in performing noncomputational activities. In his work, he forged ideas that are at least as explicative as the Euclid-like constructions and turtle geometry but more accessible and more powerful [5].

He defined CT as a problem-solving method that uses computer science techniques and concepts. Jeanette Wing [8] recently started reviving CT and emphasizing its

H. Shodiev (✉)
Wilfrid Laurier University, 75 University Ave,
Waterloo, ON N2L 3C5, Canada
e-mail: hshodiev@wlu.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_58

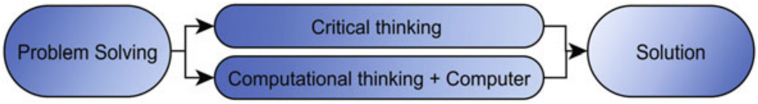
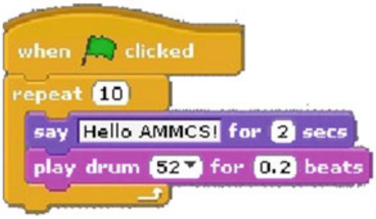


Fig. 1 2D Problem solving model

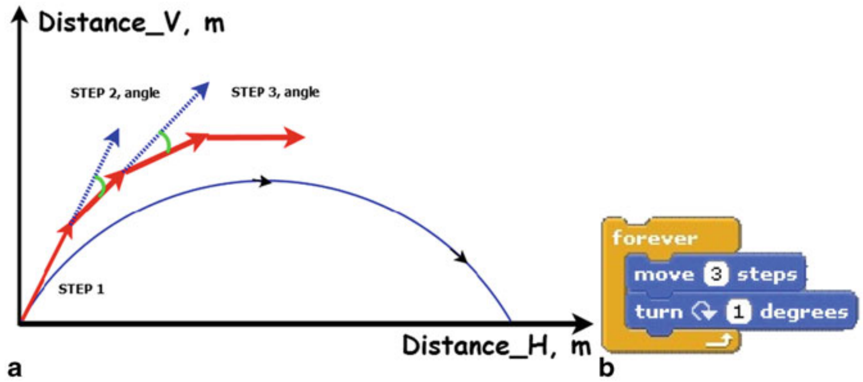


```
when clicked
repeat 10
  say Hello AMMCS! for 2 secs
  play drum 52 for 0.2 beats
```

Fig. 2 Scratch graphical blocks

role across all disciplines. She argued that CT is a fundamental skill for everyone, not just for students majoring in computer science. She initiated a profound engagement with the core questions of what computer science is and what it might contribute to solving problems across the spectrum of human inquiry. We argue that advances in educational technologies allow us to bring CT and effectively use it in secondary and postsecondary school levels. We intend to help bridge the gap between the K-12, noncomputer science disciplines and the computer science education communities by investigation of relevant age appropriate resources for science, music [6], art [1], and video games [2]. In this chapter, we propose embedding CT concepts with a universal tool called Scratch [7]. A key component to employing CT with Scratch is the possibility to visualize the phenomena which allows enhanced understanding of the concept. There is no clear evidence of using Scratch by young women. The US National Center for Women and Information Technology (NCWIT) in a case study about Scratch, calls Scratch a promising practice for increasing gender diversity in information technology [3]. Programs in Scratch can be created by simply snapping together graphical blocks, much like LEGO bricks or puzzle pieces (see Fig. 2).

There is less focus on syntax, so one is not required to add semicolons or square brackets. The blocks are designed to fit together only in ways that make sense, so there are no syntax errors as in traditional programming languages. In this study, we show an example of how Scratch can be used to simulate projectile motion in physics. A projectile is any object projected into space by the exertion of a force, i.e., a thrown basketball. In addition to projectile motion, we can simulate other physical phenomena in our immediate surroundings such as the motion of colliding spheres, conservation of momentum and others. Depending on the level of mathematics obtained in secondary school, we can create the simulation with and without algebraic tools. This simulation allows students not only to improve their CT through tinkering but also to focus on physics concepts themselves. Asking students to explain the concept can be preceded by asking them to simulate the projectile motion.



```
forever
  move 3 steps
  turn ↻ 1 degrees
```

Fig. 3 Vertical distance change due to change of the launch angle

2 Method

Simulation without complex algebra allows us to focus on results—visualization of physical phenomena such as projectile motion. Students can play with scratch code by tinkering to get parabolic trajectory. This can be done by either changing the angle to the horizontal or the vertical distance.

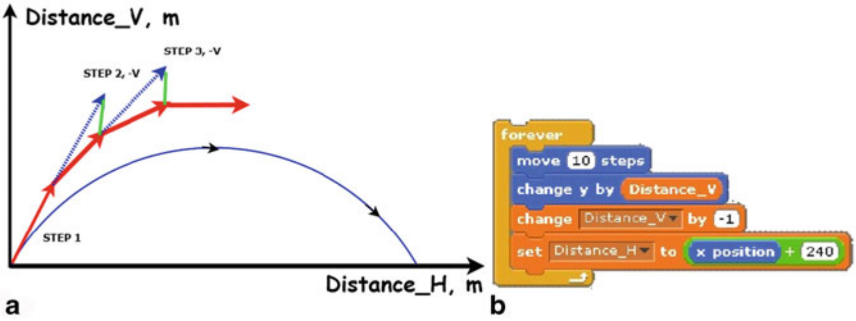
2.1 Changing the Angle of Launching

Suppose the projectile is launched in a direction defined by an initial angle with respect to the horizontal shown in Fig. 3a. This angle can be decreased by a certain amount after each iteration to reach a peak and then ultimately return to the initial *y*-position. In the example in Fig. 3b, the angle is decreased by 1° after every ten steps. Simulations for three different launch angles are shown in Fig. 3.

Vertical distance change due to change of the launch angle is

$$Y_{n+1} = Y_o - Y_\theta \tag{1}$$

so distance Y changes incrementally as the angle changes incrementally. The steps are made up of both horizontal and vertical components such that the projectile rise reaches a peak and falls with a trajectory that is symmetrical to the path toward the peak. Calculating the time of flight, the horizontal range, and the height of the projectile can be avoided at this level.



```
forever
    move 10 steps
    change y by Distance_V
    change Distance_V by -1
    set Distance_H to x position + 240
```

Fig. 4 Vertical distance change due to change of the vertical displacement

2.2 Changing the Vertical Displacement

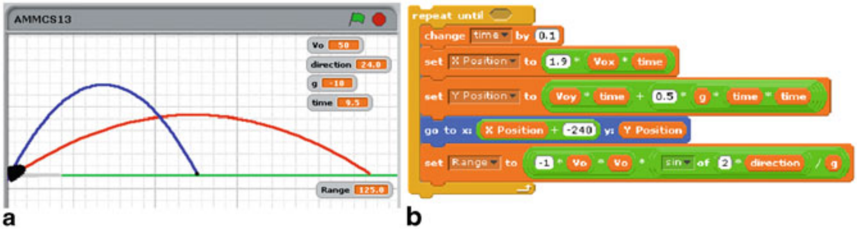
We can also visualize the trajectory by changing the vertical displacement each iteration until it reaches the initial *y*-position as shown in Fig. 4a.

In Fig. 4b, the vertical displacement of each iteration, which is made up of ten steps, is decreased by 1 units. To utilize the entire screen of the scratch interface we set the initial *x*-position to − 240 and *y*-position to − 166 as (0,0). The iteration stops once a condition of reaching the *y*-position of − 166 once again is met.

2.3 Mathematical Modeling with Algebra

At a higher level of mathematics, algebra can be applied to describe the behavior of a projectile. This can be done in terms of its kinematics motion without dealing with force or energy as a function of time. We assume that the initial and final *y*-positions of the projectile are the same. We will use the simplest example of a ball launched upward into the air at an initial angle with respect to the horizontal and velocity. In this case, we can calculate: (1) vertical displacement of the projectile at its peak, (2) horizontal displacement of the projectile, and (3) the launch angle that will result in the largest travel distance. We can then use these scalars and determine the speed and acceleration of the projectile. Visualization of projectile motion helps us to understand the dynamics of the motion. Once we visualize the motion, it becomes much easier to answer above questions and calculate them. This visualization using Scratch leads us to understand the concept of projectile motion in 2D. As a result the following benefits can be gained: (1) advanced understanding of the concept and (2) advanced problem solving skills. In this simulation, *DX* and *DY*changes along the *X*- and *Y*-axis, respectively. So using mathematics we can determine these vertical and horizontal changes at different points in time.

$$DY = Vxt \tag{2}$$



Vo 50
direction 24.0
g -10
time 9.5
Range 125.0

```
repeat until
  change time by 0.1
  set X Position to 1.9 * Vox * time
  set Y Position to Voy * time + 0.5 * g * time * time
  go to x: X Position + -240 y: Y Position
  set Range to -1 * Vo * Vo * sin of 2 * direction / g
```

Fig. 5 Vertical distance change due to change of the vertical displacement

$$DY = Vyt + \frac{1}{2}gt^2, \tag{3}$$

where Vx and Vy are x and y components of velocity, respectively. As shown, the horizontal component of velocity is constant and vertical component of velocity is varying with time. The initial velocities can be calculated by multiplying the initial velocity by the cosine or sine of the launch angle

$$Vx = V\cos\theta \tag{4}$$

$$Vy = V\sin\theta. \tag{5}$$

It is known that the projectile reaches the highest point when the Vy component of velocity is 0 m/s. The total time of travel is two times the time it takes to reach the peak vertical point. We can also determine the height H from (4) and maximum horizontal distance traveled R from Eq. (5). Scratch code for this example is shown in Fig. 5.

$$R = \frac{V^2\sin 2\theta}{g}. \tag{6}$$

3 Conclusion

Understanding basic Scratch commands and control tools can help to implement more trajectories and generative algorithms by creating and manipulating sequences of graphical commands. Using real-life science phenomena as an example, we can create innovative and interactive visualizations to tap into the imagination of students who might never have considered science as fun and playful. Students from various backgrounds tend to be intimidated by the terminology used in science. However, with more exposure to interesting projects, students can start thinking computationally and actively. CT with a hands-on scratch graphical approach gives them necessary confidence.

References

1. Brennan, K., Resnick, M.: New frameworks for studying and assessing the development of computational thinking. In: Proceedings of the 2012 Annual Meeting of the American Educational Research Association, Vancouver, Canada (2012). http://web.media.mit.edu/ mres/papers.html
2. Gee, J.P.: What Video Games Have to Teach Us About Learning and Literacy. (Second edition: revised and updated edition). Palgrave Macmillan (2007)
3. NCWIT.: What makes scratch so accessible to novices? National Center for Women & Information Technology PROMISING PRACTICES (2008)
4. Papert, S.: An exploration in the space of mathematics educations. Int. J. Comput. Math. Learn. **1**(1):95–123 (1996)
5. Papert, S., Solomon, C.: Twenty things to do with a computer. MIT AI Lab memo 248 (1971)
6. Ruthmann, A., Heines, J.M., Greher, G.R., Laidler, P., Saulters, C., II.: Teaching computational thinking through musical live coding in scratch. In: Proceedings of the 41st ACM Technical Symposium on Computer Science Education, pp. 351–355. ACM (2010)
7. Scratch graphyical platform. -http://scratch.mit.edu.
8. Wing, J.M.: Computational thinking. Commun. ACM **49**(3), 33–35 (2006)

Mathematical and Computational Modeling of Noise Characteristics of Channel Amplifiers

Alla Shymanska

Abstract This work is devoted to computational modeling of stochastic processes of the electron multiplication in electron amplifiers in order to reduce the noise factor which is a measure of the loss of available information. The effects of the processes, arising when a layer with increased secondary emission yield is formed at the entrance of the channel, are investigated.

A computational method for simulation of stochastic processes of an electron multiplication in microchannel electron amplifiers is developed. It is based on 3D Monte Carlo (MC) simulations and theorems about serial and parallel amplification stages proposed by the author. Splitting a stochastic process into a number of different stages, enables a contribution of each stage to the entire process to be easily investigated. The method provides a high calculation accuracy with minimal cost of computations. The computational model easily implements new experimental data without any changes in the algorithm.

1 Introduction

Channel electron multipliers are widely used in many areas as single devices and in microchannel plate (MCP), which is an array of single parallel channels (Fig. 1) [1, 2, 5, 7]. However, statistical fluctuations in the gain of the channels increase a noise factor which is a measure of the loss of available information. Investigations dealing with reduction of the noise factor are of considerable practical interest.

This work is devoted to the computational modeling of stochastic processes of the electron multiplication in the electron amplifiers in order to reduce the noise factor. The effects of the processes, arising when a layer with increased secondary emission is formed at the entrance of the channel, are investigated.

A. Shymanska (✉)
School of Computing and Mathematical Sciences,
Auckland University of Technology, Private Bag 92006,
Auckland 1142, New Zealand
e-mail: alla.shymanska@aut.ac.nz

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_59

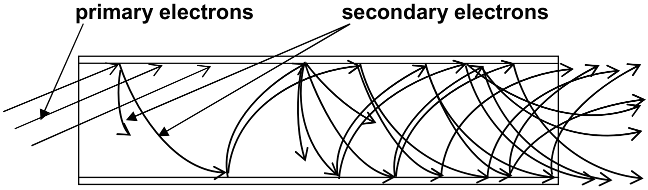


Fig. 1 Electron multiplication in the channel

The following real physical picture was considered in the modeling. The electrons of a primary parallel monochromatic beam entering the channel hit the walls at different incidence coordinates and angles, producing secondary electrons which are multiplied until they leave the channel. The secondary emission yield (SEY) of the first collision and the length along which subsequent amplification occurs in the channel are different. Therefore, amplitude distributions produced by single primary electrons are very different what increases the noise factor. Different incidence angles of the primary electrons affect SEY of the first collision and, consequently, the noise factor.

After the first collision, the primary electrons produce secondary electrons with different emission energy and directions. The secondary electrons are multiplied until they leave the channel. When all the electrons have emerged from the channel, the yield of the individual pulse is known. The gain of individual pulses is fluctuated considerably what increases the noise factor.

The high-efficiency emitter is deposited on the top of the contact conducting layer at the entrance of a channel with the purpose to increase the SEY and, therefore, to reduce the noise factor. The electrostatic field inside the channel and contact conducting layer create a nonuniform electrostatic field at the entrance of the channel, and the conditions for the movement of secondary electrons in this region are different from the motion of electrons in a uniform field inside the channel. Due to the spread in the collision coordinates, the input electrons bombard the high-efficiency emitter and the wall of the channel not coated with the high-efficiency emitter. The area covered by the input electron beam depends on the incidence angle of the primary electrons. All these factors and parameters are taken into account in the computational model developed here.

The computational method is based on 3D Monte Carlo (MC) simulations and theorems about serial and parallel amplification stages [6]. Splitting a stochastic process into a number of different stages enables a contribution of each stage to the entire process to be easily investigated. The method preserves all advantages of the MC simulations which are used only once for one simple stage. The use of the theorems allows to conduct any further investigations and optimizations without additional MC simulations. The method provides a high calculation accuracy with minimal cost of computations. The computational model easily implements new experimental data without any changes in the algorithm.

2 Computational Model

The entire multiplication process can be split into sequential stages and/or parallel multiplication paths, and how it is done depends on particular investigation. If the input signal amplification is represented as a sequence of transformations (each of which is characterized by a mean and a variance), then one can speak of serial amplification stages. If m_k is the mean and d_k is the variance of the probability distribution of the number of particles at the output of the k- stage, produced by one particle at its input, then, using logarithmic generating functions, we can obtain the mean M, and variance D of the amplitude distribution after the Nth stage [6]:

$$M = \prod_{k=0}^{N} m_k, \tag{1}$$

$$D = \sum_{k=0}^{N} d_k \prod_{i=0}^{k-1} m_i \prod_{j=k+1}^{N} m_j^2. \tag{2}$$

The expressions (1) and (2) constitute the theorem of serial amplification stages.

If the primary particle is multiplied along one of n possible parallel paths, and if each path gives an average of g_k particles at the output with a variance of v_k, then the mean G and the variance V of the amplitude distribution at the output of the system with some parallel amplification paths, can be obtained as:

$$G = \sum_{k=1}^{n} \rho_k g_k, \tag{3}$$

$$V = \sum_{k=1}^{n} \rho_k v_k + \sum_{k=1}^{n} \rho_k g_k^2 - G^2, \tag{4}$$

where ρ_k is the probability of choosing the kth path. Equations (3) and (4), constitute the theorem of parallel amplification paths [6].

For variations in the collision coordinates of the electrons of the primary beam, the portion of the channel from an elementary area at its input, where the collision occurred, to the output of the channel can be considered as the amplification path. The variance D and the average gain G at the output of the multiplier can be defined using (3) and (4), where sums should be replaced by integrals over the surface of the channel, bombarded by the electrons of the primary beam (or, correspondingly, from discrete distributions to probability densities):

$$G = \int_s \psi(s) g(s) ds, \tag{5}$$

$$D = \int_s \psi(s)d(s)ds + \int_s \psi(s)g^2(s)ds - G^2, \tag{6}$$

where s is the surface area stroked by particles, ψ is the probability density for the particle to strike the elementary surface ds, $g(s)$ is the average number of particles with variance $d(s)$ at the output of the path.

In the model, the multiplication process of a single electron, emitted at the beginning of the channel, is simulated by 3D MC methods in a homogeneous field along the effective channel length. It is defined in [6] as a part of the channel where the amplitude distribution is stabilized (the Poisson distribution at the beginning of the channel changes to the negative exponential function). The mean gain $g(z)$ and variance $d(z)$ as functions of the coordinate z along the channel axis are calculated for the single electron on the effective length, and then, using the theorems of serial and parallel amplification stages, these functions are obtained on the entire channel length.

The trajectory of each electron is calculated in three dimensions from the ballistic equations, and the position, energy, and angle of the subsequent collisions are determined. The result of each collision is calculated as before and the process is repeated for each secondary electron generated. The nonuniform field at the entrance of the channel is calculated by the finite difference method for the Laplace equation. The trajectories of the electrons in the nonuniform field are calculated by the Runge–Kutta method. The part of the channel with nonuniform field is considered as a separate stage.

The actual number of secondaries generated by the particular collision is a random sample taken from the Poisson distribution:

$$P(\nu) = \frac{\sigma^{\nu} e^{-\sigma}}{\nu}, \tag{7}$$

where ν is the number of secondary electrons produced, σ is the SEY. The variation of the SEY is defined by a secondary emission function [3].

The energy distribution is described by $p(\varepsilon) = 2.1\bar{\varepsilon}^{-3/2}\sqrt{\varepsilon}exp(-1.5\varepsilon/\bar{\varepsilon})$, where $\bar{\varepsilon}$ is the mean energy [8].

Each secondary electron is assigned two emission angles chosen from Lambert's law: $p_1(\theta) = \sin 2\theta$ and $p_2(\varphi) = 1/2\pi$, where θ is the angle between the normal to the surface and emission direction, and φ is the azimuthal angle.

The functions $g(z)$ and $d(z)$ and the theorems about serial and parallel amplification stages are used to calculate the mean G and the variance D of the distribution at the output of a channel, and thus determine the noise factor F of the channel multiplier [3, 6]:

$$F = \frac{1}{\gamma}\left(1 + \frac{D}{G^2}\right), \tag{8}$$

where γ is the fraction of the front surface of the multiplier exposed to electrons.

3 Effects of a High-Efficiency Emitter

To reduce a noise factor, a high-efficiency emitter with high SEY is deposited onto the entrance of the channels. However, in some cases it does not bring the expected result, and the noise factor can even increase [4].

If the layer with high SEY is not intact (as a result, for example, of not smooth walls inside a channel), then there would be random fluctuations of SEY along the layer of high-efficiency emitter which should affect the noise factor.

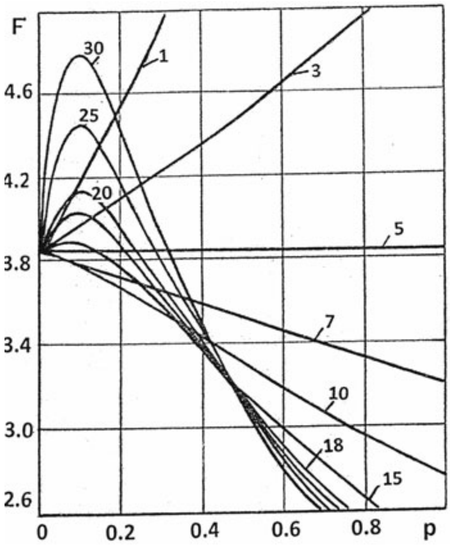
Consider two emitters with σ_1 and σ_2 randomly distributed along the layer with high SEY due to its nonuniformity, where two SEYs relate to the high-efficiency layer and the channel material. Let areas occupied by these two sections be s_1 and s_2, respectively. Probability p that primary electron enters the section with σ_1 can be defined as $\frac{s_1}{s_1+s_2}$, then the probability to enter section with σ_2 is $(1-p)$. Therefore, the probability p can be interpreted as the relative fraction of the area occupied by the emitter with average value of σ_1 when the average SEY of the rest of the surface is σ_2. Using the Eq. (7), we obtain the expression for the distribution $P(\nu)$ of the number of electrons ν knocked out by one primary electron entering a channel for the case of a nonuniform emitter:

$$P(\nu) = p\frac{\sigma_1^{\nu} e^{-\sigma_1}}{\nu} + (1-p)\frac{\sigma_2^{\nu} e^{-\sigma_2}}{\nu}, \tag{9}$$

where the mean and variance can be obtained as $m = p\sigma_1 + (1-p)\sigma_2$ and $d = p(1-p)(\sigma_1 - \sigma_2)^2 + m$ correspondingly.

Figure 2 shows the computational results of the dependence of the noise factor F on level of nonuniformity of the emitter p for different values of σ_1 (σ_2 is constant). It

Fig. 2 The dependence of the noise factor F on level of nonuniformity of the emitter p



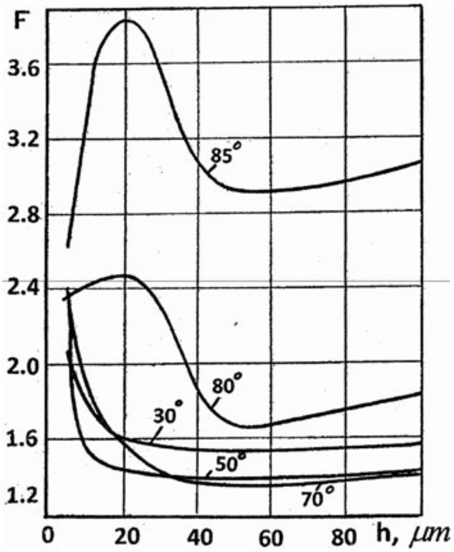


Fig. 3 The dependence of the noise factor on the length of the layer with the high secondary emission yield (SEY)

can be shown that the maximum of $F(p)$ is determined by the condition $(G_1^2 - G_2^2) > D_1 + D_2$, where G_1 and D_1 are the mean gain and the variance of the amplitude distribution, obtained with only σ_1 $(p = 1)$, and G_2 and D_2 are the mean gain and the variance with only σ_2 $(p = 0)$. Therefore, $F(p)$ has a maximum if the distributions corresponding to uniform emitters are sufficiently “far away” from one another. In this case, the spectrum of pulses, generated by emission from sections with σ_2, transforms as p increases into the spectrum with σ_1, and for some p the maximum width of the total distribution is greater than the width of each of the distributions.

Figure 3 shows the computational results of the dependence of the noise factor F on the length h of the layer with the high SEY for the different incident angles θ of the primary electron beam. The increase of the noise factor, when a high-efficiency emitter is deposited on the entrance part of a channel, is due to the “effect of the nonuniform emitter” described above. The amplitude distribution at the output of the channel is the result of a superposition of two distributions: of electrons bombarding the high-efficiency emitter and electrons bombarding the walls of the channel not coated with the high-efficiency emitter. This case is analogous to that considered earlier but here the regions with different SEYs are spatially localized. The noise factor can increase or decrease, depending on the fraction of the area of the high-efficiency emitter with respect to the area covered by the input electron beam.

It is seen from the graph that an increase of the coverage depth beyond $5d$, where $d = 10\mu m$ is the channel diameter, does not lead to a decrease in the noise factor, and for a big θ it leads even to an increase in the noise factor.

The results, obtained here, enable one to choose optimal regimes of a microchannel amplifier in terms of the noise factor.

4 Conclusions

The method for calculation of the stochastic processes has been developed which is based on two theorems of the sequential and parallel amplification stages, proved by the author.

The method have been used to describe the effects of the high-efficiency emitter on the noise factor when the spread in incidence coordinates of the primary electrons and nonuniform electrostatic field at the entrance of the channel are taken into account in the model. The conducted investigation shows the effectiveness of the method of serial and parallel stages in calculations of stochastic processes.

References

1. Choi, Y.S., Kim, J.M.: Monte Carlo simulations for tilted—channel electron multipliers. IEEE Trans. Electron Devices **47**, 1293–1296 (2000)
2. Giudicotti, L.: Analytical, steady-state model of gain saturation in channel electron multipliers. Nucl. Instr. Method A **480**, 670–679 (2002)
3. Guest, A.J.: A computer model of channel multiplier plate performance. Acta Electron **14**, 79–97 (1971)
4. Leonov, N.B., Tyutikov, A.M.: Opt.-Mekh. Prom. **2**, 10 (1983)
5. Shikhaliev, P.M., Ducote, J.L., Xu, T., Molloi, S.: Quantum efficiency of the MCP detector: Monte Carlo calculation. IEEE Trans. Nucl. Sci. **52**, 1257–1262 (2005)
6. Shymanska, A.V.: Computational modeling of stochastic processes in electron amplifiers. J. Comput. Electron **9**, 93–102 (2010)
7. Shymanska, A.V.: Numerical analysis of electron optical system with microchannel plate. J. Comput. Electron **10**, 291–299 (2011)
8. Yakobson, A.M.: Estimation of the multiplication coefficient of a secondary electron multiplier with a continuous dynode, Radiotekh. Electron **11**, 1813–1825 (1966)

Parameter Range Reduction in ODE Models in the Presence of Partial Data Sets

Andrew Skelton and Allan R. Willms

Abstract The problem of estimating parameters from time series data is considered. A parameter range reduction scheme is employed to quickly reduce a priori ranges of parameters. The effectiveness of the scheme is tested in the presence of partial data sets using an SIR model test case. The algorithm is shown to make substantial reductions of parameter ranges when limited time series data is available. Such reductions are shown to be of benefit to traditional parameter estimation techniques.

1 Introduction

We consider an ordinary differential equation model for some physical or biological process and the inverse problem of determining appropriate model parameter values from time series data. This goal is typically achieved by selecting a set of parameter values, numerically integrating the model equations and comparing the result to the time series data. A cost function such as weighted least squares gives a measure of the optimality of these parameters and new parameter values continue to be chosen until the cost function has been minimized.

It is often the case that little is known a priori about the parameter values, so making an initial guess can be difficult. If the initial choice must be made from a very large region of parameter space, it is possible that this selection will result in a system which cannot be numerically integrated over the observation time window. If the cost function has multiple local minima, or large flat regions in parameter space, it may be difficult for the procedure to converge to a reasonable minimum. To combat this problem, one could employ a multistart method (such as the methods presented in [2]) in which a large number of initial parameter choices are made to better identify the global minimum. Most of the computational time of such an approach, however, is spent numerically integrating the model, so multistart methods can add significant

A. Skelton (✉) · A. R. Willms
Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada
e-mail: skeltona@uoguelph.ca

A. R. Willms
e-mail: awillms@uoguelph.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_60

computational time. A variety of techniques have been developed to combat these problems (see [1, 5] for summaries of various methods), but all can suffer when processing large regions of parameter space.

In [8, 9], the authors developed a parameter range reduction method. If each parameter is known to lie initially within an a priori range, the algorithm quickly prunes regions of parameter space, removing boxes of parameter values that are deemed to be inconsistent with the data. The final output is a collection of consistent boxes from which a better initial parameter set can be chosen.

The original method assumed that time series data was available for all model variables. In [9], the scheme was applied to a real-world problem modelled by a five-dimensional, seven-parameter pharmacokinetic system [4, p. 152]. An algebraic equation associated with the model was used to eliminate an unmeasured variable and the parameter range reduction algorithm was applied to the resulting four-dimensional system. The algorithm was only able to make small reductions to some of the parameter ranges. We believed that more progress might be possible if the unmeasured variable was kept in the system of equations and its occurrences treated as additional parameters to be reduced. Since the true parameter values for this system are not known, we test this hypothesis using a simpler system.

2 Problem

If S, I, R denote respectively the number of susceptible, infected and recovered individuals in a population of size N, and α, γ represent the infection and recovery rates, we obtain the standard Susceptible-Infected-Recovered (SIR) model

$$S' = -\alpha S I, \tag{1}$$

$$I' = \alpha S I - \gamma I, \tag{2}$$

$$R' = \gamma I, \tag{3}$$

$$N = S + I + R. \tag{4}$$

We are interested in estimating the values of α and γ, given time series data of some subset of model variables. Data were simulated by numerically integrating Eqs. (1–3) with true parameter values $(\alpha, \gamma) = (0.05, 0.5)$ and initial conditions $(S_0, I_0, R_0) = (49, 1, 0)$. We sampled 40 equally spaced data points on the time interval $[0, 20]$. Normally distributed noise with mean 0 and standard deviation $\sigma = 0, 1, 3$ was then added to the data. Negative simulated data values were reset to half their true value.

The estimation algorithm requires all discrete time series data to be transformed to continuous representations. We used the algorithm developed in [7] to enclose the data with continuous piecewise-linear curves. The results of this banding procedure for the infected data (I) are shown in Fig. 1 for each data set. We have been overly conservative in our banding to reflect the inherent uncertainty in the data themselves.

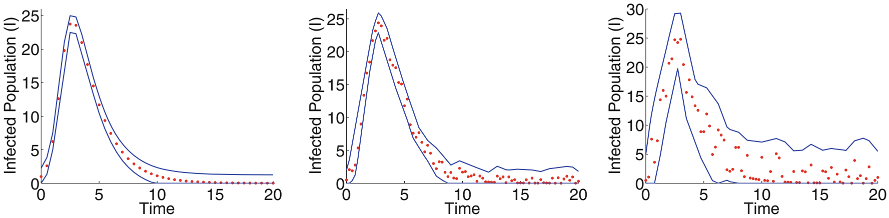


Fig. 1 Continuous bands applied to infected (I) time series data. The plots from *left* to *right* show the continuous bands applied to simulated data with noise levels $\sigma = 0, 1, 3$ respectively

3 Algorithm

The parameter range reduction algorithm is fully described in [9], so we provide only a brief overview here. For a single parameter box, the algorithm loops through discretization time windows $[t^0, t^n]$ and at each window, applies a specific family of $(n + 1)$-step linear multistep discretizations to each equation. We define a discretization of Eq. (2) to be of the form

$$F := I^0 - I^n + h \sum_{i=0}^{n} \beta_i \left(\alpha S^i I^i - \gamma I^i \right), \tag{5}$$

where S^i and I^i are approximations to the true solution at time $t^i = t^0 + ih$, and each $\beta_i \geq 0$ is chosen as described in [9]. Since each parameter (α, γ) and variable (S^i, I^i) is known to be contained in a given interval, occurrences of each quantity can be replaced by their interval-valued counterparts. Using interval arithmetic, we can obtain an *enclosure* $[\underline{\mathbf{F}}, \overline{\mathbf{F}}]$ of the true range of F over its domain. Depending on the form of the discretization equation, this enclosure interval may be wider than the true range [6], but techniques can be used to sharpen the enclosure [3].

For the given parameter intervals to be valid, the enclosure $\left[\underline{\mathbf{F}}, \overline{\mathbf{F}}\right]$ must contain 0. If it does not, then the parameter box can be discarded as inconsistent with the data. The algorithm then attempts to find regions of parameter space on which $\underline{\mathbf{F}} > 0$ or $\overline{\mathbf{F}} < 0$. Such regions can then be discarded. If no such region can be found, the algorithm splits the parameter box along its widest edge and processes the resulting boxes independently. A cap on the total number of boxes is set.

If a model variable is unmeasured, we treat each occurrence of that variable in the discretization equation as an additional parameter that can be reduced. For example, if the susceptible population is unmeasured in Eq. (5) the algorithm regards $S^0, \ldots, S^n$ as $(n+1)$ additional parameters that can also be reduced. Since arbitrary discretization windows can be chosen, a naive implementation would lead to an unacceptable number of new parameters that must be stored. We instead store unmeasured variable ranges at a fixed number of control time points. Intervals required by the discretization are then linearly interpolated from the band, and any reductions are carefully applied to the nearest control points. Details of this have been omitted from this chapter due to space considerations and will be reported elsewhere.

4 Methods and Results

For each data set, we applied the range reduction algorithm to the four-equation SIR model in Eqs. (1–4), and the equivalent model obtained by rearranging Eq. (4) to eliminate the S variable as follows:

$$I' = \alpha(N - I - R)I - \gamma I, \tag{6}$$

$$R' = \gamma I. \tag{7}$$

The parameter estimation problem for each model was considered in three scenarios. First, the case in which all model variables were measured. Second, then the case in which the infected population I and recovered population R were measured, and finally the case in which only the infected population I was measured. We tested the algorithm for all values of the step size, h, from 0.05 to 1.00 in increments of 0.05, and for all values of the discretization parameter s from 1 to 16. The values of α, γ and N were assumed to initially lie respectively in the intervals $[0.01, 200]$, $[10^{-6}, 10^{3}]$, and $[25, 75]$.

The success of the algorithm was measured as follows. For each of the parameters α and γ (not N since it is not required to simulate the original SIR model), we found the hull of all valid parameter boxes and determined the fraction of parameter space represented by this hull. The geometric mean of these two fractions, μ was used as a measure of algorithm performance. For readability, this was reported below as $-\log(\mu)$. We also computed the centre of mass of all non-discarded parameter boxes and tested the validity of this value as an initial parameter guess for a traditional parameter estimation algorithm. We used the matrix laboratory (MATLAB) function `fminsearch` with a least-squares cost function to search parameter space for a valid parameter set (α, γ and all three initial conditions). As a control, we first ran the search using the true parameter values and the midpoint of each continuous band as the initial parameter choices. We then ran the search using the centre of mass value of α and γ and the centre of the reduced box if any variables were unmeasured. If the output parameter estimate differed by not more than $0.0001, 0.001, 0.1, 0.1$ (values selected heuristically and proportionally to the true values), respectively in their approximations of α, γ, S_0 and I_0, then the approximation is deemed to be in the same basin of attraction as the true value and is highlighted in Table 1. Heuristic analysis indicated that the optimization algorithm was insensitive to the value of R_0 and a poor estimate did not appear to affect the results.

When time series data was available for all three model variables, using the four-equation model allowed the algorithm to make more significant reductions in parameter space. For almost all cases, the centre of mass approximation was an excellent starting point for a traditional minimization routine. When the susceptible population was unmeasured, it can be seen that the four-equation model still outperforms the reduced model. Thus, it appears that treating unmeasured variables as additional parameters is a better strategy than attempting to remove such quantities from the model. It would appear that the increased algebraic complexity of the equations in the reduced model is a serious hindrance to parameter range reduction.

Table 1 Results. The algorithm outputs a set of consistent boxes. We calculate the hull and centre of mass of this set of boxes for each parameter and unmeasured initial conditions. We also report $-\log(\mu)$, a measure of the size of the hull relative to the initial size of parameter space

Noise	Model	Parameter Hull		Centre of Mass			$-\log(\mu)$
		α	γ	(α,γ)	S_0	R_0	
S,I,R all measured							
$\sigma = 0$	2	$[0.0321, 0.0740]$	$[0.427, 0.587]$	$(0.0515, 0.509)$	-	-	3.740
	4	$[0.0416, 0.0632]$	$[0.437, 0.572]$	$(0.0526, 0.504)$	-	-	3.918
$\sigma = 1$	2	$[0.0236, 0.0969]$	$[0.385, 0.623]$	$(0.0562, 0.5061)$	-	-	3.530
	4	$[0.0327, 0.0716]$	$[0.403, 0.597]$	$(0.0534, 0.496)$	-	-	3.712
$\sigma = 3$	2	$[0.0100, 0.223]$	$[0.133, 1.021]$	$(0.104, 0.603)$	-	-	3.014
	4	$[0.0129, 0.143]$	$[0.218, 0.910]$	$(0.0787, 0.562)$	-	-	3.173
I,R measured, S unmeasured							
$\sigma = 0$	2	$[0.0321, 0.0740]$	$[0.427, 0.587]$	$(0.0515, 0.509)$	-	-	3.740
	4	$[0.0322, 0.0656]$	$[0.431, 0.580]$	$(0.0499, 0.508)$	49.498	-	3.801
$\sigma = 1$	2	$[0.0236, 0.0969]$	$[0.385, 0.623]$	$(0.0562, 0.5061)$	-	-	3.530
	4	$[0.0250, 0.0827]$	$[0.386, 0.602]$	$(0.0529, 0.498)$	49.281	-	3.602
$\sigma = 3$	2	$[0.0100, 0.223]$	$[0.133, 1.021]$	$(0.104, 0.603)$	-	-	3.014
	4	$[0.0100, 0.137]$	$[0.220, 0.909]$	$(0.0676, 0.580)$	50.198	-	3.180
I measured, S,R unmeasured							
$\sigma = 0$	2	$[0.010, 200]$	$[0.000, 0.699]$	$(109, 0.325)$	–	30.050	1.578
	4	$[0.0105, 0.166]$	$[0.319, 0.840]$	$(0.102, 0.449)$	32.058	21.930	3.197
$\sigma = 1$	2	$[0.010, 200]$	$[0.000, 0.729]$	$(110, 0.336)$	–	30.050	1.569
	4	$[0.018, 0.110]$	$[0.121, 1.005]$	$(0.0707, 0.4093)$	29.537	24.783	3.195
$\sigma = 3$	2	$[0.010, 200]$	$[0.000, 1.184]$	$(110, 0.599)$	–	30.050	1.463
	4	$[0.010, 0.102]$	$[0.000, 1.740]$	$(0.0610, 0.703)$	28.426	27.544	3.048

This complexity affects the monotonicity of the equation which results in decreased sharpness in the interval computations. In the case when only the infected population is measured, the four-equation model dramatically outperforms the reduced model. It is worth noting, however, that neither model was able to immediately provide a good enough starting guess for the minimization routine. We were, however, able to significantly reduce the amount of parameter space a traditional algorithm would be required to search.

This procedure is computationally very fast. When all three variables are measured, the algorithm typically finished in approximately 0.1 clock seconds. When only the infected population was measured, the algorithm typically finished in less than 0.5 clock seconds. Simulations were conducted with a 2.4-GHz Intel Core i5 processor.

When the algorithm failed to provide a good starting value, it was often due to the poor estimation of the initial conditions. In this chapter, while the unmeasured variables were treated as parameters in the context of the reduction scheme, to avoid a combinatorial explosion of boxes, these unmeasured variable intervals were not allowed to be split. When we use the four-equation model with no-noise data in the scenario in which only the infected population is measured, allowing all unmeasured variable intervals to be split into five subintervals at each time step allows us to obtain the vastly improved starting estimate $S_0 = 42.294$ and $R_0 = 6.564$. This improvement, however, causes the run time of the algorithm to increase from 0.38 clock seconds to 38.55 clock seconds. We are currently working on a method by which these ranges can be split without adding significant computational time.

5 Conclusion

In this chapter, we tested the parameter range reduction algorithm on a standard SIR model. We found that the scheme worked better on models with a larger number of simpler equations, even if time series data was unavailable for some model variables. We found that in most cases, the algorithm was able to substantially reduce the size of the valid parameter box and in many cases was also able to produce an excellent starting point for a traditional parameter estimation scheme. The considerably smaller parameter box provides a much smaller region for a global minimizer to search, thus providing a significant computational savings.

Acknowledgement This work was supported by an Ontario Graduate Scholarship, a Natural Sciences and Engineering Council of Canada (NSERC) Postgraduate Scholarship and an NSERC Discovery Grant.

References

1. Bard, Y.: Comparison of gradient methods for the solution of nonlinear parameter estimation problems. SIAM J. Numer. Anal. **7**(1), 157–186 (1970)
2. Guus, C., Boender, E., Romeijn, H.: Stochastic methods. In: Horst, R., Pardalos, P.M. (eds.) Handbook of Global Optimization. Kluwer Academic Publishers, Dordrecht, Netherlands. 829869 (1995)
3. Hansen, E., Walster, G.W.: Global Optimization Using Interval Analysis. M. Dekker, New York (2004)
4. Knott, G.D., Kerner, D.R.: MLAB Applications Manual. Civilized Software Inc., Bethesda (2004)
5. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. **13,** 2467–2474 (2003)
6. Moore, R.E.: Interval Analysis. Prentice-Hall, Englewood Cliffs (1966)
7. Skelton, A., Willms, A.R.: An algorithm for continuous piecewise linear bounding of discrete time series data. BIT Numer. Math. **54**(4), 1155–1169 (2014)
8. Willms, A.R.: Parameter range reduction for ode models using cumulative backward differentiation formulas. J. Comput. Appl. Math. **203**(1), 87102 (2007)
9. Willms, A.R., Szusz. E.K.: Parameter range reduction for ODE models using monotonic discretizations. J. Comput. Appl. Math. **247,** 124–151 (2013)

Stabilization of Impulsive Systems via Open-Loop Switched Control

Peter Stechlinski and Xinzhi Liu

Abstract In this chapter, the stabilization of nonlinear impulsive systems under time-dependent switching control is investigated. In the open-loop approach, the switching rule is programmed in advance and the switched system is composed entirely of unstable subsystems. Sufficient conditions are found that establish the existence of stabilizing time-dependent switching rules using the Campbell–Baker–Hausdorff formula and Lyapunov stability theory.

1 Introduction

Recently there has been increased interest in switched systems, which are systems governed by a combination of continuous/discrete dynamics and logic-based switching and have applications in many real-world problems (for example, see [1–3] and the references therein). The current literature on the stabilization of unstable continuous systems using switching control can be separated into two categories: the first is the closed-loop switched control problem, which involves the construction of a stabilizing state-dependent switching rule (first studied by Wicks et al. in [4] and further in, for example, [5–8]). The second is the open-loop switched control problem, which has been studied much less extensively, and entails the construction of a stabilizing time-dependent switching rule calculated a priori and hence preprogrammed into the data. The authors Bacciotti and Mazzi [9] studied nonlinear switched systems and found sufficient conditions for the existence of a solution to the open-loop problem. Stabilization of nonlinear systems to a compact set using a time-dependent switching rule was considered by Mancilla-Aguilar and Garcia in [10]. In [11], Bacciotti and Mazzi investigated eventually periodic switching rules for the linear switched problem.

P. Stechlinski (✉) · X. Liu
University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: pstechli@uwaterloo.ca

X. Liu
e-mail: xzliu@uwaterloo.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_61

Impulsive control was considered in [12, 13] for the state-dependent switching rule approach; however, to the best of the authors' knowledge there has been no work done on the open-loop approach with impulses present. Hence, the objective of the present report is to extend the current literature by considering stabilization of nonlinear impulsive systems using high-frequency time-dependent switching. The main contributions are verifiable conditions for the existence of a stabilizing time-dependent switching rule for both disturbance impulses and stabilizing impulses. The rest of the chapter is outlined as follows: in Sect. 2, the open-loop switched control problem with impulses is presented. Then, in Sect. 3, sufficient conditions are proved which guarantee the existence of a stabilizing time-dependent switching rule. An example is given in Sect. 4 and some discussions done in Sect. 5.

2 Problem Formulation

Let $\mathbb{R}^n$ denote the Euclidean space of n-dimensions equipped with the Euclidean norm $\|\cdot\|$ and let $D \subset \mathbb{R}^n$ be an open set. Let $\mathbb{R}_+$ denote the set of nonnegative real numbers. Consider the following class of functions for later use:

$$\mathcal{K} = \{w \in C(\mathbb{R}_+, \mathbb{R}_+) : w \text{ is strictly increasing and } w(0) = 0\}.$$

Consider the following switched impulsive system

$$\begin{cases} \dot{x} = f_\sigma(x), & t \neq \tau_k, \\ \Delta x = g_k(x), & t = \tau_k, \\ x(0) = x_0, & k = 1, 2, \ldots \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state vector; $\sigma : (t_{k-1}, t_k] \to \{1, 2, \ldots, m\}$, where m is a positive integer greater than one, is the switching rule with switching instances t_k that satisfy $0 < t_1 < t_2 < \ldots < t_{k-1} < t_k < \ldots$ with $t_k \to \infty$ as $k \to \infty$; $\{f_i\}_{i=1}^m$ is a family of sufficiently smooth functions that satisfy $f_i : D \to \mathbb{R}^n$ and $f_i(0) = 0$ for $i = 1, \ldots, m$. Here, $\Delta x := x(t^+) - x(t)$ and $x(t^+) := \lim_{a \to 0^+} x(t + a)$ and the impulsive moments τ_k satisfy $0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k < \ldots$ with $\tau_k \to \infty$ as $k \to \infty$. The impulsive functions $\{g_k\}_{k=1}^\infty$ are continuous and satisfy $x + g_k(x) \in D$ for all $x \in D$. Note that system (1) can be derived from a control system perspective (see, for example, [12, 14]).

The goal of the open-loop switched control problem is as follows: given a set of vector fields $\{f_i\}_{i=1}^m$ such that each subsystem $\dot{x} = f_i(x)$ is unstable, stabilizing or disturbance impulses $\{g_k\}_{k=1}^\infty$ with associated impulsive moments $\{\tau_k\}_{k=1}^\infty$, and initial condition x_0, find a time-dependent switching rule $\sigma(t)$ a priori such that the trivial solution of (1) is asymptotically stable.

3 Main Results

In order to give the main results a lemma is first required, which follows from the Campbell–Baker–Hausdorff formula (see [15]).

Lemma 1 [9]
Let $\mathcal{H}$ be the space of all bounded, analytic, vector fields on the open ball $B_r(0)$ (for some constant $r > 0$), equipped with an appropriate norm so that it is a Banach space. Suppose that $f_1, f_2 \in \mathcal{H}$ and let α_1, α_2 be positive constants such that $\alpha_1 + \alpha_2 = 1$. Then there exists a positive constant $\bar{T}$ such that for all $T < \bar{T}$ and for all $x \in B_r(0)$, the series

$$h(x) =(\alpha_1 f_1(x) + \alpha_2 f_2(x))T + \frac{\alpha_1\alpha_2}{2}T^2[f_1, f_2](x) + \ldots + c_n(x) + \ldots \quad (2)$$

converges where $[f_1, f_2](x) := (df_1)_x f_2(x) - (df_2)_x f_1(x)$ denotes the Lie product, df_x is the Jacobian matrix of the vector field f, and the n^{th} term in the series $c_n(x) = c_n(x; T, \alpha_1, \alpha_2)$ is defined recursively.[1] *Furthermore, $\varphi_1(\alpha_1 T, \varphi_2(\alpha_2 T, x)) = \varphi_h(1, x)$, where φ_h is the flow generated by the vector field h.*

We are now in a position to present the first stability result, which considers stabilizing switching control and impulsive disturbances. The theorem (and proof) is an extension of the work in [9] and is based on the existence of a stable convex combination of the subsystems.

Theorem 1 *Assume that there exist constants $\alpha_i > 0$ for $i = 1, \ldots, m$ satisfying $\sum_{i=1}^{m} \alpha_i = 1$, constants $\lambda, a_k, \epsilon_k > 0$, $0 < \delta_k < 1$, and functions $w_1, w_2 \in \mathcal{K}$, and $V \in C^1[\mathbb{R}^n, \mathbb{R}_+]$ such that for $k = 1, 2, \ldots,$*

(i) $w_1(\|x\|) \leq V(x) \leq w_2(\|x\|)$ *for all* $x \in D$;
(ii) $\dot{V} \leq -\lambda V$ *along solutions of* $\dot{x} = \sum_{i=1}^{m} \alpha_i f_i(x)$;
(iii) $V(x + g_k(x)) \leq (1 + a_k)V(x)$ *for all* $x \in D$;
(iv) $\ln(1 + a_k) - (1 - \epsilon_k)(\tau_k - \tau_{k-1})\lambda < \ln \delta_k$.

Then there exists a time-dependent switching rule, possibly dependent on the initial condition, such that the trivial solution of (1) is asymptotically stable.

Proof Consider the case $m = 2$ (to extend the proof to $m > 2$, see the comments in [9]). For any constant $c > 0$, define L_i to be the connected component of the level set $\{x : V(x) < c \prod_{j=0}^{i} \delta_j\}$ where $\delta_0 = 1$. Without loss of generality, consider a value of c and $r > 0$ such that: the closure of L_0, denoted $cl(L_0)$, is contained in the basin of attraction of $x = 0$; $cl(L_0) \subset B_r(0) \subset D$; and $x_0 \in cl(L_0) \setminus L_1$.

By Lemma 1 there exists a constant $T_0 > 0$ such that $h(x)$ in Eq. 2 converges for all $T \in [0, T_0]$ and $x \in cl(L_0)$. Then, along solutions of $\dot{x} = h(x)$,

[1] See Chap. 2 in [15] for the details.

$$
\begin{aligned}
\dot{V} &= \nabla V \cdot [(\alpha_1 f_1(x) + \alpha_2 f_2(x))T + \Big(\frac{\alpha_1\alpha_2}{2}[f_1, f_2](x)\Big) T^2 + \ldots],\\
&= \nabla V \cdot T \sum_{i=1}^{2} \alpha_i f_i(x) + \nabla V \cdot \Big[\Big(\frac{\alpha_1\alpha_2}{2}[f_1, f_2](x)\Big) T^2 + \ldots\Big],\\
&\leq -\lambda V T + R(x)T^2,
\end{aligned}
$$

where $R(x)$ is a continuous function. Define the closed set $\mathcal{V}_0 := cl(L_0) \setminus L_1$ and choose $T' > 0$ sufficiently small so that

$$\left|\max_{x\in\mathcal{V}_0} R(x)\right| T' \leq \epsilon_1 \lambda \min_{x\in\mathcal{V}_0} w_1(\|x\|).$$

Thus, $\dot{V} \leq -\lambda V T + \epsilon_1 \lambda V T = -\lambda(1-\epsilon_1)VT$ for all $T \in [0, \min\{T_0, T'\}]$ and $x \in \mathcal{V}_0$. In particular, $\dot{V} \leq -\lambda(1-\epsilon_1)VT_1$ for $x \in \mathcal{V}_0$, where T_1 is chosen so that $0 < T_1 \leq \min\{T_0, T'\}$ and $mod(\tau_1, T_1) = 0$. Let η_1 be the positive integer such that $\eta_1 T_1 = \tau_1$ then $V(\varphi_h(\eta_1; x_0)) \leq V_0 e^{-(1-\epsilon_1)\lambda\eta_1 T_1}$. Define $\Phi_{T_1}(x) := \varphi_2(\alpha_2 T_1, \varphi_1(\alpha_1 T_1, x))$ and let $\Phi_{T_1}^{\eta_1}(x)$ indicate η_1 iterations of the mapping, that is, $\Phi_{T_1}(\Phi_{T_1}(\ldots(x))\ldots)$ (η_1 times). Then, by Lemma 1, $V(\Phi_{T_1}^{\eta_1}(x_0)) \leq V_0 e^{-(1-\epsilon_1)\lambda\eta_1 T_1}$. Hence,

$$V(x(\eta_1 T_1; x_0)) \leq V_0 e^{-(1-\epsilon_1)\lambda\eta_1 T_1}$$

along solution trajectories of (1).

An impulse is applied to the switched system state trajectory at $\tau_1 = \eta_1 T_1$:

$$V(x(\eta_1 T_1^+; x_0)) \leq (1+a_1)V(x(\eta_1 T_1; x_0)) \leq (1+a_1)V_0 e^{-(1-\epsilon_1)\lambda\eta_1 T_1}.$$

Condition (iv) implies that $V(x(\eta_1 T_1^+; x_0)) < \delta_1 V_0 < \delta_1 c$ for all $x_0 \in \mathcal{V}_0$. The first segment of the switching rule can then be constructed as

$$\sigma(t; x_0) = \begin{cases} 1, & t \in (jT_1, jT_1 + \alpha_1 T_1],\ j = 0, 1, \ldots, \eta_1 - 1, \\ 2, & t \in (jT_1 + \alpha_1 T_1, (j+1)T_1]. \end{cases}$$

Set $x_1 = x(\eta_1 T_1^+; x_0)$ then $x_1 \in \mathcal{V}_1 := cl(L_1) \setminus L_2$ and by the above arguments it can be shown that there exist a positive constant T_2 and a positive integer η_2 such that $\eta_2 T_2 = \tau_2 - \tau_1$ and $V(x(\eta_2 T_2^+; x_1)) < \delta_1\delta_2 c$ for all $x_1 \in \mathcal{V}_1$. By repeating the process,

$$V(x(\eta_k T_k^+; x_{k-1})) < \delta_1\delta_2\cdots\delta_k c$$

for all $x_{k-1} \in \mathcal{V}_{k-1} := cl(L_{k-1}) \setminus L_k$ for some constant $T_k > 0$ and positive integer η_k satisfying $\eta_k T_k = \tau_k - \tau_{k-1}$. The result follows. □

For stabilizing impulses, the following result can be applied.

Theorem 2 *Assume that there exist constants $\alpha_i > 0$ for $i = 1, \ldots, m$ satisfying $\sum_{i=1}^m \alpha_i = 1$, constants $\lambda, a_k, \epsilon_k > 0$, $0 < \delta_k < 1$, and functions $w_1, w_2 \in \mathcal{K}$, and $V \in C^1[\mathbb{R}^n, \mathbb{R}_+]$ such that for $k = 1, 2, \ldots,$*

(i) $w_1(\|x\|) \leq V(x) \leq w_2(\|x\|)$ *for all* $x \in D$*;*
(ii) $\dot{V} \leq \lambda V$ *along solutions of* $\dot{x} = \sum_{i=1}^{m} \alpha_i f_i(x)$*;*
(iii) $V(x + g_k(x)) \leq a_k V(x)$ *for all* $x \in D$*;*
(iv) $\ln a_k + (1 + \epsilon_k)(\tau_k - \tau_{k-1})\lambda < \ln \delta_k$.

Then there exists a time-dependent switching rule, possibly dependent on the initial condition, such that the trivial solution of (1) *is asymptotically stable.*

Proof The proof is similar to the proof of Theorem 1.

4 Example

Consider system (1) with $m = 2$, impulsive moments $\tau_k = 2k$ for $k = 1, 2, \ldots,$

$$f_1(x_1, x_2) = \begin{pmatrix} 5x_1 + 2x_2^5 - x_2^2 e^{sinx_1} \\ -3x_2 - 2x_1 x_2^4 \end{pmatrix}, \quad f_2(x_1, x_2) = \begin{pmatrix} -6x_1 - x_2^5 \\ 2x_2 + x_1 x_2^4 + x_1 x_2 e^{sinx_1} \end{pmatrix},$$

$$g_{2k}(x_1, x_2) = \begin{pmatrix} sin(x_1)\sqrt{\left(1 + \frac{1}{e^{2k}}\right)(x_1^2 + x_2^2)} - x_1 \\ cos(x_1)\sqrt{\left(1 + \frac{1}{e^{2k}}\right)(x_1^2 + x_2^2)} - x_2 \end{pmatrix}, \quad g_{2k-1}(x_1, x_2) = \begin{pmatrix} 0.224\ x_1 \\ 0.224\ x_2 \end{pmatrix}.$$

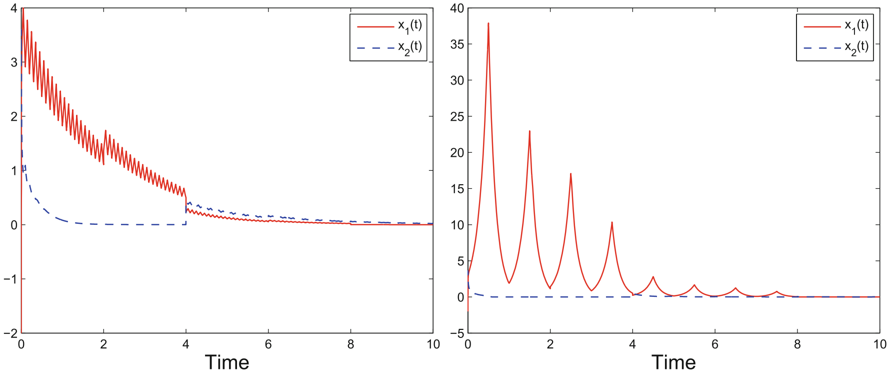
Note that both $df_{1_x}(0)$ and $df_{2_x}(0)$ have eigenvalues with positive real part. Take $\alpha_1 = \alpha_2 = 0.5$ then

$$\sum_{i=1}^{2} \alpha_i f_i(x_1, x_2) = \frac{1}{2} \begin{pmatrix} -x_1 + x_2^5 - x_2^2 e^{sinx_1} \\ -x_2 - x_1 x_2^4 + x_1 x_2 e^{sinx_1} \end{pmatrix}.$$

Consider the Lyapunov function $V = x_1^2 + x_2^2$, then along $\dot{x} = \alpha_1 f_1(x) + \alpha_2 f_2(x)$, $\dot{V} = -(x_1^2 + x_2^2) = -V$. At the impulsive times, $x_1(\tau_{2k}^+)^2 + x_2(\tau_{2k}^+)^2 \leq (1 + \frac{1}{e^{2k}})(x_1(\tau_{2k})^2 + x_2(\tau_{2k})^2)$ and $x_1(\tau_{2k-1}^+)^2 + x_2(\tau_{2k-1}^+)^2 \leq 1.5(x_1(\tau_{2k-1})^2 + x_2(\tau_{2k-1})^2)$. Let $w_1(\|x\|) = w_2(\|x\|) = \|x\|^2$, $\lambda = 1$, $\epsilon_{2k} = \epsilon_{2k-1} = 0.01$, $a_{2k} = 1/e^2$, and $a_{2k-1} = 0.5$. Then, the conditions of Theorem 1 are satisfied with $\delta_k = 0.5$ and hence there exists a stabilizing time-dependent switching rule. From the simulations (see Fig. 1), it is apparent that the origin is asymptotically stable if the systems are switched every 0.05 time units (periodic switching rule with $T = 0.1$) or 0.5 time units ($T = 1$).

5 Conclusions

Although motivated from a control problem, the analysis in the present chapter is more general since the impulses can be considered as disturbances or can be stabilizing forces. In both cases, we have given sufficient conditions for the existence



$T = 0.1$ (198 total switches).
a

$T = 1$ (18 total switches).
b

Fig. 1 Simulation of (1) with $x_0 = (-2, 3)$

of a stabilizing time-dependent switching rule. There are many reasons why a time-dependent switching rule may be desired over a state-dependent one [9]: for example, with a time-dependent approach, chattering, sliding motions, and Zeno behaviours can be avoided (see [1, 8]). Since the time-dependent switching rule is preprogrammed into the data, sensors are not as vital. There are some drawbacks to this approach: currently there is no explicit formula for the time-dependent switching rule (a possible direction for future work). The number of switches required in this approach might be unrealistic physically (see the example above).

Acknowledgments This research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Liberzon, D.: Switching in Systems and Control. Birkhauser, Boston (2003)
2. Liberzon, D., Morse, A.S.: Basic problems in stability and design of switched systems. IEEE Control Syst. Mag. **19**(5), 59–70 (1999)
3. Shorten, R., Wirth, F., Mason, O., Wulff, K., King, C.: Stability criteria for switched and hybrid systems. SIAM Rev. **49**(4), 545–592 (2007)
4. Wicks, M., Peleties, P., DeCarlo, R.: Switched controller synthesis for the quadratic stabilization of a pair of unstable linear systems. Eur. J. Control **4**(2), 140–147 (1998)
5. Kim, S., Campbell, S., Liu, X.: Stability of a class of linear switching systems with time delay. IEEE Trans. Circuits Syst. I: Regul. Paper **53**(2), 384–393 (2006)
6. Gao, F., Zhong, S., Gao, X.: Delay-dependent stability of a type of linear switching systems with discrete and distributed time delays. Appl. Math. Comput. **196**(1), 24–39 (2008)
7. Hien, L., Ha, Q., Phat, V.: Stability and stabilization of switched linear dynamic systems with time delay and uncertainties. Appl. Math. Comput. **210**(1), 223–231 (2009)

8. Liu, J., Liu, X., Xie, X.: On the (h0,h)-stabilization of switched nonlinear systems via state-dependent switching rule. Appl. Math. Comput. **217**(5), 2067–2083 (2010)
9. Bacciotti, A., Mazzi, L.: Stabilisability of nonlinear systems by means of time-dependent switching rules. Int. J. Control **83**(4), 810–815 (2010)
10. Mancilla-Aguilar, J.L., García, R.A.: Some results on the stabilization of switched systems. Automatica **49**(2), 441–447 (2013)
11. Bacciotti, A., Mazzi, L.: Asymptotic controllability by means of eventually periodic switching rules. SIAM J. Control Optim. **49**(2), 476–497 (2011)
12. Liu, X., Stechlinski, P.: Hybrid control of impulsive systems with distributed delays. Nonlinear Anal.: Hybrid Syst. **11**, 57–70 (2014)
13. Wang, Q., Liu, X.: Stability criteria of a class of nonlinear impulsive switching systems with time-varying delays. J. Franklin Inst. **349**(3), 1030–1047 (2012)
14. Guan, Z.H., Hill, D., Shen, X.: On hybrid impulsive and switching systems and application to nonlinear control. IEEE Trans. Autom. Control **50**(7), 1058–1062 (2005)
15. Varadarajan, V.: Lie Groups, Lie Algebras, and Their Representations. Springer-Verlag, New York (1984)

Mathematics-in-Industry Study Group Projects from Australia and New Zealand in the Past Decade

Winston L. Sweatman

Abstract Mathematics in Industry Study Groups (MISG) have been an annual event in Australia and New Zealand since 1984. Projects from the last decade are considered. Among the industries involved are those of steel, electricity and agriculture.

1 Introduction

The Mathematics in Industry Study Group (MISG) workshops in Australia and New Zealand were initiated by Australia's national science agency (CSIRO) in 1984. At present, the workshops occur annually as a special interest group meeting of the Australia and New Zealand Industrial and Applied Mathematics (ANZIAM) organisation. During the last decade the workshops have been hosted in turn by Massey University, New Zealand (2004, 2005, 2006), University of Wollongong, Australia (2007, 2008, 2009), Royal Melbourne Institute of Technology (RMIT) University, Australia (2010, 2011, 2012) and Queensland University of Technology (QUT), Australia (2013) [1].

Each workshop lasts for one week (in late January or early February). The week begins with presentations from industry representatives, during which they describe their project. Thereafter, small teams of participants work on each individual project led by two (or sometimes three) moderators. Continuing interaction and discussion with the industry representatives helps to further formulate and make progress with the project. As well as coordinating their group, the moderators are responsible for reporting on progress during presentations mid-week and at the end of the week, and afterwards in written reports published after the conclusion of the workshop.

During the 10-year period (2004–2013), each of the workshops involved between 4 and 7 industry projects. A total of 57 projects were considered in the decade. Several of the industrial partners returned to the workshops on multiple occasions.

W. L. Sweatman (✉)
Centre for Mathematics in Industry, Institute of Natural and Mathematical Sciences,
Massey University, Albany, Auckland, New Zealand
e-mail: w.sweatman@massey.ac.nz

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_62

The projects have been varied. Some of the kinds of projects tackled and industrial partners involved are:

- Steel: New Zealand Steel and Bluescope Steel Research
- Electric power: Transpower, Integral Energy
- Whiteware: Fisher and Paykel
- Agriculture: Plant Protection Chemistry New Zealand, Fonterra, NRM/Tegel, Compac Sorting Equipment
- Ecological: Environment Canterbury, Australian Institute of Marine Science
- Medical: Brain Research Institute, Kirby Institute
- Others: Geoscience Australia, Australian Bureau of Statistics, Defence Science and Technology Organisation, Department of Transport and Main Roads

In the following sections, some of the projects are described in more detail. The author was a MISG team member on these ten projects and was also a moderator for them all except for the first (in 2004).

2 New Zealand MISG Projects 2004, 2005, 2006

The MISG 2004–2006 were the first to be based in New Zealand. Six or seven projects were brought in each year.

Modelling of a Poultry Shed: NRM/Tegal Ltd., 2004 This related to the large barns in which chickens are raised for meat over a 6-week period. During this time, water and food is fed to the birds. The chickens themselves produce moisture and heat, which are removed from the shed by ventilation. The project involved modelling this flow of mass and energy. Among diverting considerations were the appropriate surface areas of chickens when standing or seated (spheres or hemispheres) [2].

Implementing Lanier's Patents: Backyard Technology, 2005 Behind this project was the idea that aeroplanes would fly better with appropriate 'holes' in their wings that Lanier expressed in the 1930s. Unfortunately, there was no data to work with, the closest thing being a few patents and photographs from the 1930s. This made things rather challenging. However, the MISG team did assess and summarise what information there was in the patents, and also conducted simple analyses and numerical simulations relating to the conjectures [3].

Process Driven Models for Spray Retention of Plants: Plant Protection Chemistry NZ, 2006 Modelling the deposit of horticultural sprays onto plant leaves can be helpful for designing the implementation to be more effective. A number of processes can be modelled separately such as the transit of the spray from the spray nozzle to the plant, the impact of an individual droplet on a leaf and the flow of droplets across leaves. In the second of these processes, the droplet may leave the leaf through the two different mechanisms of rebound and shatter (possibly to return to the leaf for additional impacts) or the droplet may remain on the leaf. The group collected these processes together to form a composite model of the process [4]. A continuation of research after the MISG considered the leaf impaction in more detail [5].

Further to the project at MISG 2006, Plant Protection Chemistry NZ were involved in the MISG in 2005 and 2013, these projects involving the passage of spray through porous barriers (hedges) and the uptake of agrichemicals through leaves.

3 Electric Power Projects

During the decade, Transpower Ltd., who manage the New Zealand electricity network, brought seven projects to the workshops. The projects included ones relating to maintaining electricity supply, electricity price structures, and issues relating to wind power. Integral Energy also brought electric power projects.

Operating and Planning an Electricity Transmission Grid to Maximise the Contribution of Wind: Transpower/EECA NZ, 2007 Increasing use of wind power brings new challenges because of its variable nature: at some times there may not be any wind and at other times the wind may be too powerful to be safe to use. In this project the group considered two issues. One related to how to ensure electricity supply with other power sources, when utilising a large amount of intermittent wind power. The other considered the allowances that require to be made to provide sufficient line capacity, when supply is moving between wind power and other power sources. Both studies considered the project in the context of both the financial and electric grid used in New Zealand [6, 7].

How Far can a Simplified Network Rights Auction be Extended?: Transpower, 2012 Financial options are to be introduced to complement the existing pricing structure for electricity at different locations in New Zealand. These are sold by auction and enable the purchaser to buy power at one point in the electricity network and to be supplied with it at another point in the network. The initial scheme will operate on a subset of the nodes already used for spot pricing. The MISG team considered the kind and number of constraints involved in ensuring a workable system and the feasibility of their computation. Further, an approach was suggested for generating the feasible set [8].

4 Steel Projects

Eight projects were brought to the MISGs by New Zealand Steel and Bluescope Steel Research. These considered the steel-making production from initial processes with raw iron through to the final products.

Cold Point Determination in Heat Treated Steel Coils: New Zealand Steel, 2008 Annealling is required following the production of steel sheets by rolling. This reforms the crystalline structure by a period of heating in a furnace. The steel sheets are in the form of cylindrical coils. The team considered the process of heating within these rolls and the location of the point that takes the longest to heat. It is difficult to measure temperatures within a furnace. An initial stage of the modelling was to decide upon appropriate boundary conditions for the coil which is heated by a mixture of conduction, convection and radiation. Also, within a coil, the conduction of heat is not isotropic because of the gaps between layers of the steel sheet. A series solution was found for the partial differential equations that describe the coil temperatures [9–11].

Coating Deformation in the Jet Stripping Process: Bluescope Steel Research, 2009 Steel sheets are galvanised by passing the sheet through a bath of molten coating and then controlling the thickness with air knives. Recent changes in the coating mixture has led to some potential issues to tackle with the quality of the final coating. There is a potential problem with deformations in the coating in the form of pock marks and the like. The process had been mathematically modelled previously [12]. This model was recovered but with the addition of a term due to shear stress. Numerical models indicated how potential deformations may grow [13–15].

Recovery of Vanadium During Steel Manufacturing: New Zealand Steel, 2011 In New Zealand, raw iron is produced from iron sand. The molten iron contains a number of metalloids including vanadium. These must be removed before the steelmaking process and these are also valuable by-products. The removal is done by oxidising the metalloids using oxygen blown into the molten iron and added solid iron oxides. The metalloid oxides rise to the surface of the raw iron from whence they can be scraped into another vessel. The MISG team built up a representative set of differential equations to describe the constituent substances present and the temperature. Special care must be taken of the residual carbon both as this is required later for making steel and because carbon oxidation can be a runaway process leading to carbon boil in which the molten iron is splattered everywhere [16].

5 Further Projects

Two other projects were moderated by the author. Both of the organisations: the Australian Defence Science and Technology Organisation (DSTO) and Fonterra Cooperative Group Ltd. (Fonterra) supported the workshops in multiple years.

Influence Diagrams to Support Decision Making: DSTO, 2010 The 2010 MISG team considered influence diagrams. These can be used in a variety of ways, they indicate links (arrows) between events or actions (boxes) as a support for decision making. The team spent a great deal of time exploring the possibilities for these approaches in their discussion. They visualised the approach being used in a hierarchical fashion where detail in sub-influence diagrams could be hidden until required within larger networks. A computational simulator was created that used colour shades to show the state of different events and this was helpful for visualising the progress of influence through a system [17].

Can we Predict How Cheese Matures?: Fonterra, 2013 Cheddar cheese is sampled soon after production before storage for ripening. The MISG team produced a differential equations model for key processes involved including ones for the breakdown and consumption of proteins, fats and carbohydrates by bacteria and enzyme-catalysed reactions. Data from the literature was fitted with this simple model. Further to this, data from the Fonterra factory was analysed. The evolution of acidity in the process was also modelled [18].

6 Concluding Remarks

The MISG over the last decade have tackled varied projects. Key points of a selection of these have been presented. The workshops continue to be a productive and instructive venture for participants from both industry and academia.

Acknowledgement The author is grateful to the many MISG participants over the decade: Directors, Industry Representatives, Moderators and Group Members. It has been instructive and enjoyable!

References

1. The Mathematics in Industry Study Group (MISG) internet site, Queensland University of Technology (QUT): http://mathsinindustry.com/ (2014). Accessed 17 May 2014
2. McKibbin, R., Wilkins, A.: Modelling of a poultry shed. In: Wake, G.C. (ed.) Proceedings of the 2004 Mathematics-in-Industry Study Group, pp. 47–59 (2005)
3. Hocking, G.C., Stokes, Y.M., Sweatman, W.L.: Implementing Lanier's patents for stable, safe and economical ultra-short wing Vacu- and Para-planes. In Wake, G.C. (ed.) Proceedings of the 2005 Mathematics-in-Industry Study Group, pp. 119–141 (2005)
4. Mercer, G., Sweatman, W.L., Elvin, A., Caunce, J., Fulford, G., Harper, S., Pennifold, R.: Process driven models for spray retention by plants. In: Wake, G.C. (ed.) Proceedings of the 2006 Mathematics-in-Industry Study Group, pp. 57–85 (2007)
5. Mercer, G.N., Sweatman, W.L., Forster, W.A.: A model for spray droplet adhesion, bounce or shatter at a crop leaf surface. In: Fitt, A.D., Norbury, J., Ockendon, H., Wilson, E. (eds.) Progress in Industrial Mathematics at ECMI 2008 (Mathematics in Industry 15), pp. 937–943. Springer, Berlin (2010)
6. Pritchard, G., Sweatman, W.L., Nan, K., Camden, M., Whiten, W.: Maximizing the contribution of wind power in an electric power grid. In: Merchant, T., Edwards, M., Mercer, G. (eds.) Proceedings of the 2007 Mathematics and Statistics in Industry Study Group, pp. 114–139 (2008)
7. Sweatman, W.L., Pritchard, G., Whiten, W., Camden, M., Nan, K.: Optimising for wind power contributions in an electricity grid. In: Fitt, A.D., Norbury, J., Ockendon, H., Wilson, E. (eds.) Progress in Industrial Mathematics at ECMI 2008, Mathematics in Industry 15, Springer Berlin Heidelberg, pp. 1031–1037 (2010)

8. Pritchard, G., Sweatman, W.L., Mohammadian, G., Kilby, P.: A simplified financial transmission rights auction in the context of the New Zealand electricity grid. ANZIAM J. **54**, M83–M104 (2013)
9. McGuinness, M., Sweatman, W.L., Baowan, D., Barry, S.I.: Annealing steel coils. In: Merchant, T., Edwards, M., Mercer, G. (eds.) Proceedings of the 2008 Mathematics and Statistics in Industry Study Group, pp. 61–80 (2009)
10. Barry, S.I., Sweatman, W.L.: Modelling heat transfer in steel coils. ANZIAM J. **50**, C668–C681 (2009)
11. Sweatman, W.L., Barry, S.I., McGuinness, M.: Heat transfer during annealing of steel coils. In: Günther, M., Bartel, A., Brunk, M., Schöps, S., Striebel, M. (eds.) Progress in Industrial Mathematics at ECMI 2010 (Mathematics in Industry 17), pp. 303–309. Springer, Berlin 2012)
12. Tuck, E.O.: Continuous coating with gravity and jet stripping. Phys. Fluids **26**, 2352–2358 (1983)
13. Hocking, G.C., Sweatman, W.L., Roberts, M.E., Fitt, A.D.: Coating Deformations in the continuous hot-dipped galvanizing process. In: Merchant, T., Edwards, M., Mercer, G. (eds.) Proceedings of the 2009 Mathematics and Statistics in Industry Study Group, pp. 75–89 (2010)
14. Hocking, G.C., Sweatman, W.L., Fitt, A.D., Breward, C.: Deformations arising during air-knife stripping in the galvanisation of steel. In: Günther, M., Bartel, A., Brunk, M., Schöps, S., Striebel, M. (eds.) Progress in Industrial Mathematics at ECMI 2010 (Mathematics in Industry 17), pp. 311–317. Springer, Berlin (2012)
15. Hocking, G.C., Sweatman, W.L., Fitt, A.D., Breward, C.: Deformations during jet-stripping in the galvanizing process. J. Eng. Math. **70**, 297–306 (2011)
16. Sweatman, W.L., Wake, G.C., Fullard, L., Bruna, M.: Recovering vanadium during the production of steel from iron sand. ANZIAM J. **53**, M1–M21 (2012)
17. Sweatman, W.L., Wake, G.C., Cooper, H.: Using influence diagrams as a tool for decision making. ANZIAM J. **52**, M147–M170 (2011)
18. Sweatman, W.L., Psaltis, S., Dargaville, S., Fitt, A., Gibb, A., Lawson, B., Marion, K.: The mathematical modelling of cheese ripening. ANZIAM J. **55**, M1–M38 (2014)

Symmetric Four-Body Problems

Winston L. Sweatman

Abstract The gravitational N-body problem has long been a source of theoretical investigation with application to astronomical systems. There is a rich and varied dynamics. With systems of four bodies arranged symmetrically, the symmetry tends to reduce the complexity of the system so that it is perhaps more similar to one with three bodies, although such systems also provide a starting point for our understanding of more general four-body systems.

1 Introduction

The gravitational N-body problem, where a number of point masses move under a mutual force between them, is of interest both as a dynamical system and because of its application to astronomy. If we take the number of bodies to be just four, then the system may be taken to represent the interaction of two binary objects, perhaps two stars with accompanying planets or alternatively a pair of binary stars.

N-body systems that begin with a symmetric set of masses, positions, and velocities, remain symmetrical for all time. This enables us to consider the symmetrical systems as a subclass of the full N-body systems.

The appeal in considering a gravitational system which is symmetric is partially due to its greater simplicity than a general N-body system. It can, however, provide insight and ideas. Some important orbits in the general problem are symmetric and can initially be found and studied more readily in a symmetric context. Symmetric orbits can be rather beautiful!

W. L. Sweatman (✉)
Centre for Mathematics in Industry, Institute of Natural and Mathematical Sciences,
Massey University, Albany, Auckland, New Zealand
e-mail: w.sweatman@massey.ac.nz

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_63

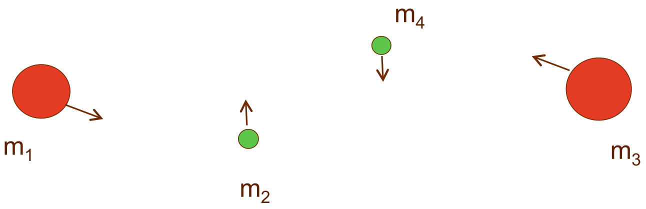


Fig. 1 The Caledonian four-body problem

2 The Caledonian Four-Body Problem

In Fig. 1, a simple symmetric four-body system in the plane is presented. The system has a rotational symmetry about its centre of mass. Mass m_3 is equal to and the image of mass m_1, and mass m_4 is equal to and the image of mass m_2. Relative to the centre of mass, which we fix at the origin, the state of the system is fully described by the positions and momenta of the two masses m_1 and m_2. This system has been named the Caledonian four-body problem [4, 5, 9, 10]. Another system with a similar complexity can be obtained by using a reflective rather than rotational symmetry. Further, if two perpendicular lines of symmetry and four identical bodies are used, then a system results whose state is captured by the position of a single mass.

3 The Symmetrical Collinear Four-Body Problem

Apart from symmetry, another simplification occurs by considering N-body systems in a lower dimensional space. In particular, the Caledonian four-body problem reduces to the symmetrical four-body problem (Fig. 2) when the orbits of all masses lie on a fixed line [11, 12]. If the initial positions and velocities are all collinear then again, as with symmetry, the orbit will retain this special property for all time. A further property is that the order of the masses on the line does not change.

If we denote the (one dimensional) positions of masses m_1 and m_2 relative to the centre of mass by x_1 and x_2, respectively, their conjugate momenta are $w_1 = 2m_1\dot{x}_1$ and $w_2 = 2m_2\dot{x}_2$ and the Hamiltonian for the symmetrical four-body system is given by

$$H = \frac{1}{4m_1}w_1^2 + \frac{1}{4m_2}w_2^2 - \frac{m_1^2}{2x_1} - \frac{m_2^2}{2x_2} - \frac{2m_1m_2}{x_1 + x_2} - \frac{2m_1m_2}{x_1 - x_2}. \tag{1}$$

A singularity will occur whenever two or more masses collide and such collisions are inevitable in the collinear problem. However, collisions between pairs of masses can be readily regularised. As eccentricity tends to unity, the natural limiting orbit of a planar encounter between a pair of masses is an elastic bounce. For the numerical

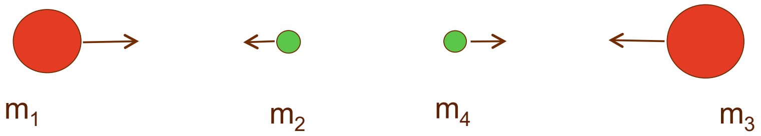


Fig. 2 The collinear four-body problem

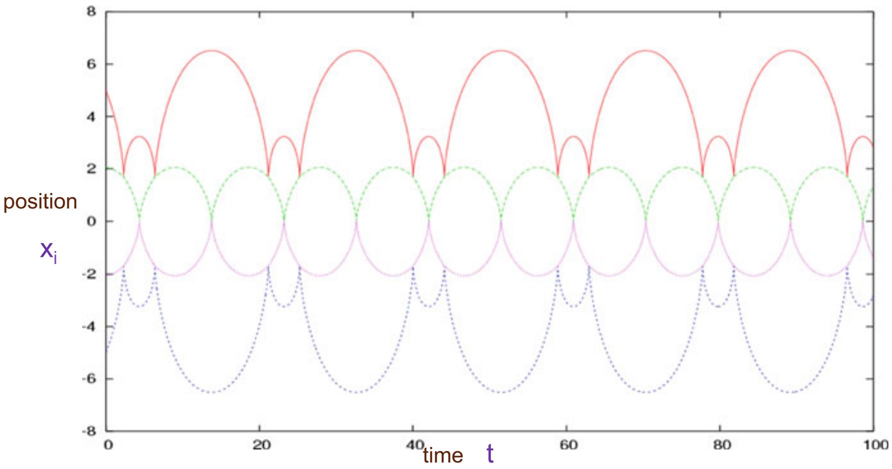


Fig. 3 A collinear four-body orbit. The different lines represent different masses

integration scheme, such encounters can be regularised by changing position coordinates to ones corresponding to the square-root of the inter-mass distances, using conjugate momenta, and an appropriate rescaling of time [11, 12]. A similar regularisation also works for removing pairwise collision singularities in the Caledonian four-body problem [8]. In general, collisions involving more than two masses lead to singularities that cannot be regularised.

Figure 3 illustrates a symmetrical collinear four-body orbit. This particular portion of the orbit is quite regular. The outer masses in this case are relatively small compared with the central masses (they are smaller by a factor of just over 26). As a result the motion of the central masses is primarily a binary motion about one another. The outer masses are orbiting this binary.

The total energy of the illustrated system (Fig. 3) is negative. Any encounter between four masses with a positive energy must result in prompt scattering of the component masses. In these circumstances, the four masses cannot even be temporarily bound. With a negative total energy, most orbits can be categorised into three types. In some orbits the masses come together in essentially a single encounter involving all the masses and spend the rest of time apart as subsystems. In other orbits there are multiple encounters involving all the masses which temporarily form subsystems between the encounters and permanently separate into subsystems after all the encounters are finished. In the third type, all the masses remain bound together for all time in periodic and quasiperiodic orbits. Various symmetrical collinear four-body orbits, in particular for the equal masses case and the Schubart-like orbits, have been presented previously [11, 12].

4 Initial Conditions and Poincaré Section Surfaces

The state of a symmetrical collinear four-body system is characterised by four parameters representing the positions and momenta of masses m_1 and m_2. However, if we fix energy $E = -1$, then the orbits lie in three-dimensional space and, by choosing a specific point or time on each orbit, the orbits can be parameterised by just two quantities within a Poincaré section [11, 12]. To define this surface we require the ratio of the position coordinates (x_2/x_1) take a fixed value α. A good choice for α is the value which corresponds to the homothetic orbit (in which orbit the masses remain in an invariant configuration leading to a quadruple collision). The surface can be parametrised by coordinates relating to position and momentum which we shall denote ρ and θ. These are based on coordinates used in the similar case of the collinear three-body problem [2, 3]. (An alternative but equivalent parameterisation for the equal masses case has been given in [7].) The position coordinate, ρ, is the separation of the outer bodies as a fraction of their maximum possible separation while on the surface $x_2/x_1 = \alpha$. The momentum coordinate, θ, chosen to be independent of R, satisfies the implicit equation

$$\tan(\theta) = \frac{\sqrt{m_1 m_2}\,(\dot{x}_2 - \alpha\dot{x}_1)}{\alpha m_2 \dot{x}_2 + m_1 \dot{x}_1}. \tag{2}$$

The value of θ is taken to be within $[0°, 180°)$ for $\dot{x}_1 > 0$ (i.e., the outer masses moving apart) and within $[180°, 360°)$ for $\dot{x}_1 < 0$. Momentum coordinates differing by 180° relate to the same orbit but with the directions of the velocities reversed. The values $\theta = 0°$ and $\theta = 180°$ correspond to the homothetic orbit which is, respectively, expanding and (then) contracting.

Such surfaces have been used for studying the equal masses case [7, 11]. Poincaré sections are formed by the crossing points of the surface by orbits from a grid of initial conditions. Alternatively a set of initial conditions may be taken within the surface and shaded according to the nature of the orbit.

Figure 4 shows a Poincaré section for a different case to the previous studies. In this case, the outer masses are nine times larger than the inner ones. This figure is one of a large number of such surfaces generated for the complete range of masses, they will be reported on in a subsequent publication. Figure 4 shows the typical structure of such sections. The regions containing the circular structures correspond to bound four-body systems. In the centre of these circles is a regular periodic orbit analogous to the collinear three-body orbit found by Schubart [6] and extended to the general-mass three-body case by Hénon [1]. Such orbits have also been found for the general-mass symmetrical four-body family [12]. Around these regions lies another region

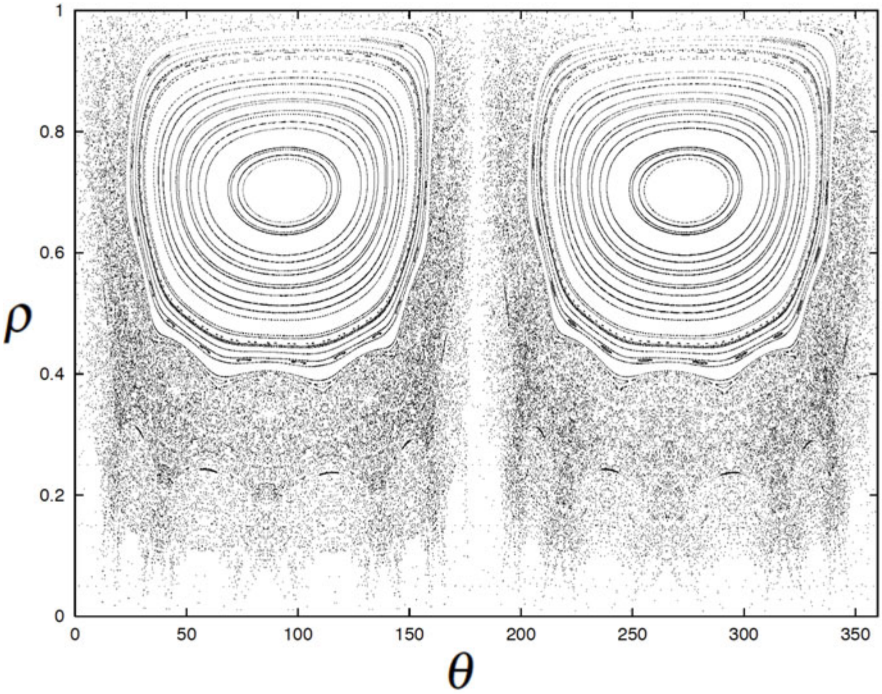


Fig. 4 A Poincaré section for which the outer masses are nine times larger than the inner masses

where the four masses remain temporarily bound undergoing a multiple number of four-body encounters before eventually separating. The systems that rapidly separate after a single encounter occur in the arches (or scallops) at the base of the figure.

5 Concluding Remarks

The four-body problem has been considered for the special case where there is symmetry. A rotational symmetry about the centre of mass leads to the Caledonian four-body problem, a system parameterised by just two position coordinates and their corresponding momenta. A further restriction to one dimension leads to the symmetrical collinear four-body problem. In this latter case the system can be described by just four variables. The orbits for this case are readily viewed as graphs of the positions of the masses against time. With a fixed value of the total energy, a Poincaré section has been presented which can be used to study the two-dimensional families of orbits.

References

1. Hénon, M.: Stability of interplay motions. Celest. Mech. Dyn. Astron. **15,** 243 (1977)
2. Hietarinta, J., Mikkola, S.: Chaos in the one-dimensional gravitational three-body problem. Chaos **3,** 183 (1993)
3. Mikkola, S., Hietarinta, J.: A numerical investigation of the one-dimensional Newtonian three-body problem III. Celest. Mech. Dyn. Astron. **51,** 379 (1991)
4. Roy, A.E., Steves, B.A.: Some special restricted four-body problems-II: from Caledonian to Copenhagen. Planet. Space Sci. **46,** 1475 (1998)
5. Roy, A.E., Steves, B.A.: The Caledonian symmetrical double binary four-body problem: surfaces of zero velocity using the energy integral. Celest. Mech. Dyn. Astron. **78,** 299 (2001)
6. Schubart, J.: Numerische Aufscuchung periodischer Lösungen im Dreikörperproblem. Astron. Nachr. **283,** 17 (1956)
7. Sekiguchi, M., Tanikawa, K.: On the symmetric collinear four-body problem. Publ. Astron. Soc. Jpn. **56,** 235 (2004)
8. Sivasankaran, A., Steves, B.A., Sweatman, W.L.: A global regularisation for integrating the Caledonian symmetric four-body problem. Celest. Mech. Dyn. Astron. **107,** 157 (2010)
9. Steves, B.A., Roy, A.E.: Some special restricted four-body problems-I: modelling the Caledonian problem. Planet. Space Sci. **46,** 1465 (1998)
10. Steves, B.A., Roy, A.E.: Surfaces of separation in the Caledonian symmetrical double binary four-body problem. In: Steves, B.A., Maciejewski, A.J. (eds.) The Restless Universe: Application of Gravitational N-body Dynamics to Planetary, Stellar and Galactic Systems. 301. IOP Publishing, Bristol (2001)
11. Sweatman, W.L.: The symmetrical one-dimensional Newtonian four-body problem: a numerical investigation. Celest. Mech. Dyn. Astron. **82,** 179 (2002)
12. Sweatman, W.L.: A family of symmetrical Schubart-like interplay orbits and their stability in the one-dimensional four-body problem. Celest. Mech. Dyn. Astron. **94,** 37 (2006)

A Simple Method for Quasilinearity Analysis of DAEs

Guangning Tan, Nedialko S. Nedialkov and John D. Pryce

Abstract We present a simple method for quasilinearity (QL) analysis of differential-algebraic equations (DAEs). It uses the signature matrix and offsets computed by Pryce's structural analysis and determines if a DAE is QL in its leading derivatives. Our method is suitable for an implementation through operator overloading or source code translation.

1 Introduction

We are interested in solving initial value problems in DAEs of the general form

$$f_i(t, \text{ the } x_j \text{ and derivatives of them}) = 0, \quad i = 1, \ldots, n, \tag{1}$$

where the $x_j(t)$, $j = 1, \ldots, n$ are state variables, and t is the time variable.

Based on Pryce's structural analysis (SA) [4], we solve (1) numerically using Taylor series, as implemented in the DAETS solver [2]. On each integration step, we compute Taylor coefficients for the solution up to some order, where we solve systems of equations for these coefficients in stages. Up to stage zero, a system can be linear or nonlinear in the variables being solved for, and after this stage, the systems are always linear.

We present a simple method for deciding if such a system is linear in the unknown derivatives, respectively Taylor coefficients. We refer to such systems as quasilinear (QL). If the unknowns appear nonlinearly, we have a nonquasilinear (NQL) system. Such information is used to determine what solver to use and the minimum number of variables and derivatives of them that need initial conditions; for details see [5].

G. Tan (✉) · N. S. Nedialkov
Department of Computing and Software, McMaster University, Hamilton, ON, Canada
e-mail: tang4@mcmaster.ca

N. S. Nedialkov
e-mail: nedialk@mcmaster.ca

J. D. Pryce
Cardiff School of Mathematics, Cardiff University, Cardiff, UK
e-mail: prycejd1@Cardiff.ac.uk

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_64

Section 2 summarizes Pryce's SA. Sect. 3 gives the definitions needed for our method. It is described in Sect. 4 and illustrated in an example in Sect. 5. Conclusions are given in Sect. 6.

2 Summary of Pryce's SA

This SA [4] constructs for (1) an $n \times n$ *signature matrix* $\Sigma = (\sigma_{ij})$ such that

$$\sigma_{ij} = \begin{cases} \text{the highest order of the derivative to which } x_j \text{ occurs in } f_i\text{; or} \\ -\infty \text{ if } x_j \text{ does not occur in } f_i. \end{cases}$$

A highest value transversal (HVT) is a set of n positions (i, j) with one entry in each row and each column, such that the sum of these entries is maximized over all transversals. From Σ, we find a HVT and equation and variable offsets $\mathbf{c}$ and $\mathbf{d}$, respectively, which are non-negative integer n-vectors satisfying

$$d_j - c_i \geq \sigma_{ij} \quad \text{for all } i, j \text{ with equality on an HVT.}$$

When the SA succeeds [1, 4], using these offsets, we can determine structural index (which is an upper bound for the differentiation index, and often they are the same), degrees of freedom, and a solution scheme for computing derivatives of the solution.

They are computed in stages $k = k_d, k_d + 1, \ldots, $ where $k_d = -\max_j d_j$. Denote

$$x_{J_k} = \left\{ x_j^{(d_j+k)} \mid d_j + k \geq 0 \right\}, \quad x_{J_{<k}} = \left\{ x_j^{(r)} \mid d_j + k > r \geq 0 \right\}, \quad \text{and}$$
$$f_{I_k} = \left\{ f_i^{(c_i+k)} \mid c_i + k \geq 0 \right\}.$$

At stage k, we solve a system of equations $f_{I_k}(t, x_{J_{<k}}, x_{J_k}) = 0$ for x_{J_k}, where $x_{J_{<k}}$ are computed at earlier stages. A system at stage $k = k_d, k_d + 1, \ldots, 0$ can be QL or NQL, while for stages $k > 0$ the systems are always linear.

Example 1 We show below for the simple pendulum (PEND), an index-3 DAE, the signature matrix and offsets. (The state variables are $x, y,$ and λ; G is gravity, and $L > 0$ is the length of the pendulum.) There are two HVTs, marked with $\bullet$ and $*$, respectively.

$$\begin{aligned} 0 &= f_1 = x'' + x\lambda \\ 0 &= f_2 = y'' + y\lambda - G \\ 0 &= f_3 = x^2 + y^2 - L^2 \end{aligned} \quad \longrightarrow \quad \Sigma = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ d_j \end{array} \begin{array}{c} \begin{array}{ccc} x & y & \lambda \end{array} \\ \left[\begin{array}{ccc} 2^{\bullet} & & 0^{*} \\ & 2^{*} & 0^{\bullet} \\ 0^{*} & 0^{\bullet} & \end{array}\right] \\ \begin{array}{ccc} 2 & 2 & 0 \end{array} \end{array} \begin{array}{c} c_i \\ 0 \\ 0 \\ 2 \\ \\ \end{array}$$

The equations for stages $k = -2, -1, 0$ are

k	$f_{I_k}(t, x_{J_{<k}}, x_{J_k})$	$x_{J_{<k}}$	x_{J_k}	Linearity
-2	$f_3 = x^2 + y^2 - L^2$	$-$	x, y	NQL
-1	$f_3' = 2xx' + 2yy'$	x, y	x', y'	QL
0	$f_1 = x'' + x\lambda$ $f_2 = y'' + y\lambda - G$ $f_3'' = 2(xx'' + x'^2 + yy'' + y'^2)$	x, x', y, y'	x'', y'', λ	QL

Obviously, at $k = -2$ we have a NQL problem, and then two QL problems.

3 Quasilinearity at Stage *k*

Definition 1 The system

$$f_{I_k}(t, x_{J_{<k}}, x_{J_k}) = 0 \tag{2}$$

is QL, if x_{J_k} appears linearly in it, and NQL otherwise.

Definition 2 A DAE is QL, if at stage $k = 0$, (2) is QL, and NQL otherwise.

Definition 3 Equation i at stage k is QL, if $f_i^{(k+c_i)} = 0$ is linear in the x_{J_k} occurring in it, and NQL otherwise.

If $c_i + k > 0$, then $f_i^{(k+c_i)} = 0$ is always QL. For example, in PEND at stage $k = -1$, $f_3' = 2xx' + 2yy' = 0$ is QL in x' and y'.

At stage k, consider equations i for which $c_i + k = 0$. If each such $f_i = 0$ is QL, then (2) is QL. If at least one such $f_i = 0$ is NQL, then (2) is NQL; cf. in PEND at stage $k = -2$.

Therefore, to determine quasilinearity at stage k, we need to check for QL only the $f_i = 0$ for which $c_i + k = 0$.

4 Algorithm

For simplicity in our exposition, we consider the code list for evaluating the f_i's as consisting of assignment, unary, and binary operators. This is the case when executing the function for evaluating the DAE through operator overloading. Our algorithm consists of initialization and propagation of *offset* and *type* data through the code list as described below.

Initialization We derive from Σ the $n \times n$ *offset* matrix $\Theta = (\theta_{ij})$ as

$$\theta_{ij} = \begin{cases} \sigma_{ij} & \text{if } \sigma_{ij} = d_j - c_i \\ +\infty & \text{otherwise,} \end{cases}$$

and derive from Θ the $n \times n$ *type* matrix $\mathtt{T} = (\mathtt{T}_{ij})$ as

$$\mathtt{T}_{ij} = \begin{cases} \mathtt{L} & \text{(Linear)} & \text{if } \theta_{ij} = 0 \\ \mathtt{U} & \text{(Undetermined)} & \text{if } 0 < \theta_{ij} < +\infty \\ \mathtt{C} & \text{(Constant)} & \text{if } \theta_{ij} = +\infty. \end{cases}$$

Then we associate with each x_j an *offset vector* $\gamma(x_j)$ being the jth column of Θ, and a *type vector* $\mathtt{T}(x_j)$ being the jth column of $\mathtt{T}$.

Propagation We propagate these vectors through the code list of the DAE according to the following rules.

R1. If $v = +u$ or $v = -u$, then

$$\gamma(v) = \gamma(u) \quad \text{and} \quad \mathtt{T}(v) = \mathtt{T}(u).$$

R2. If $v = g(u)$ is nonlinear, then

$$\gamma(v) = \gamma(u) \quad \text{and} \quad \mathtt{T}_i(v) = \begin{cases} \mathtt{N} \quad \text{(Nonlinear)} & \text{if } \mathtt{T}_i(u) = \mathtt{L} \\ \mathtt{T}_i(u) & \text{otherwise.} \end{cases}$$

R3. If $w = g(u, v)$, then for all $i = 1, \dots n$,

$$\gamma_i(w) = \min\{\gamma_i(u), \gamma_i(v)\} \quad \text{and}$$

$$\mathtt{T}_i(w) = \begin{cases} \mathtt{N} & \text{if } \mathtt{T}_i(u) = \mathtt{T}_i(v) = \mathtt{L} \,\&\, g \text{ nonlinear} \\ \max\{\mathtt{T}_i(u), \mathtt{T}_i(v)\} & \text{otherwise.} \end{cases}$$

Here we use the ordering

$$\mathtt{C} < \mathtt{U} < \mathtt{L} < \mathtt{N}.$$

R4. Consider $w = g(u, v)$. If u is a constant or the time variable t while v is not, then

$$\gamma(w) = \gamma(v) \quad \text{and} \quad \mathtt{T}(w) = \mathtt{T}(v).$$

Similarly, if v is a constant or the time variable t while u is not, then

$$\gamma(w) = \gamma(u) \quad \text{and} \quad \mathtt{T}(w) = \mathtt{T}(u).$$

R5. If $v = d^p u/dt^p$ (where $p > 0$), then

$$\gamma_i(v) = \gamma_i(u) - p \quad \text{and} \quad \mathtt{T}_i(v) = \begin{cases} \mathtt{L} & \text{if } \gamma_i(v) = 0 \\ \mathtt{U} & \text{if } 0 < \gamma_i(v) < +\infty \\ \mathtt{C} & \text{if } \gamma_i(v) = +\infty. \end{cases}$$

After executing the code list for an f_i and using R1–R5, we conclude that $f_i = 0$ is QL at stage $k = -c_i$ if $\mathtt{T}_i(f_i) = \mathtt{L}$, and NQL if $\mathtt{T}_i(f_i) = \mathtt{N}$.

5 Example

We illustrate the above method on the following index-7 DAE

$$
\begin{aligned}
0 &= f_1 = x'' + x\lambda \\
0 &= f_2 = y'' + y\lambda + (x')^3 - G \\
0 &= f_3 = x^2 + y^2 - L^2 \\
0 &= f_4 = u'' + u\mu \\
0 &= f_5 = (w''')^2 + w\mu - G \\
0 &= f_6 = u^2 + w^2 - (L + c\lambda)^2 + \lambda''
\end{aligned}
\tag{3}
$$

derived from a two-coupled pendula problem, an index-5 DAE, with originally

$$f_2 = y'' + y\lambda - G, \quad f_5 = w'' + w\mu - G, \quad f_6 = u^2 + w^2 - (L + c\lambda)^2.$$

State variables are $x, y, \lambda, u, w,$ and μ; G is gravity, $L > 0$ is the length of the first pendulum, and $c > 0$ is a constant.

We wish to determine if the DAE (3) is QL; that is, if (3) is QL at stage $k = 0$. The corresponding matrices are

$$
\begin{array}{c|cccccc|c}
 & x & y & \lambda & u & w & \mu & c_i \\
\hline
f_1 & 2 & & 0 & & & & 4 \\
f_2 & 1 & 2 & 0 & & & & 4 \\
f_3 & 0 & 0 & & & & & 6 \\
f_4 & & & & 2 & & 0 & 0 \\
f_5 & & & & & 3 & 0 & 0 \\
f_6 & & & 2 & 0 & 0 & & 2 \\
\hline
d_j & 6 & 6 & 4 & 2 & 3 & 0 &
\end{array}
\qquad
\begin{array}{c|cccccc|c}
 & x & y & \lambda & u & w & \mu & c_i \\
\hline
 & 2 & & 0 & & & & 4 \\
 & & 2 & 0 & & & & 4 \\
 & 0 & 0 & & & & & 6 \\
 & & & & 2 & & 0 & 0 \\
 & & & & & 3 & 0 & 0 \\
 & & & 2 & 0 & & & 2 \\
\hline
d_j & 6 & 6 & 4 & 2 & 3 & 0 &
\end{array}
\qquad
\begin{array}{cccccc}
x & y & \lambda & u & w & \mu \\
\hline
\mathrm{U} & & \mathrm{L} & & & \\
 & \mathrm{U} & \mathrm{L} & & & \\
\mathrm{L} & \mathrm{L} & & & & \\
 & & & \mathrm{U} & & \mathrm{L} \\
 & & & & \mathrm{U} & \mathrm{L} \\
 & & \mathrm{U} & \mathrm{L} & &
\end{array}
.
$$

Σ, blanks denote $-\infty$ $\qquad$ Θ, blanks denote $+\infty$ $\qquad$ T, blanks denote C

Note that $d_1 - c_2 > \sigma_{2,1}$ and $d_5 - c_6 > \sigma_{6,5}$; hence these σ's do not appear in Θ.

Since $c_i = 0$ for Eqs. 4 and 5, we need to examine only these two equations. (The remaining $f_i^{(c_i+k)} = 0$ are QL since $c_i > 0$ for $i = 1, 2, 3, 6$.)

Consider $f_5 = (w''')^2 + w\mu - G = 0$ with unknowns w''' and μ. We initialize

$$\gamma_5(w) = 3,\ \mathtt{T}_5(u) = \mathtt{U} \quad \text{and} \quad \gamma_5(\mu) = 0,\ \mathtt{T}_5(\mu) = \mathtt{L},$$

and propagate.

	Code list	Evaluates	$\gamma_5(v)$	$\mathtt{T}_5(v)$	Applying
	$v_4 = \mathrm{Dif}(w, 3)$	$= w'''$	0	L	R5
	$v_5 = v_4^2$	$= (w''')^2$	0	N	R2
	$v_6 = w * \mu$	$= w\mu$	0	L	R3
	$v_7 = v_5 + v_6$	$= (w''')^2 + w\mu$	0	N	R3
$f_5 =$	$v_8 = v_7 - G$	$= (w''')^2 + w\mu - G$	0	N	R4

Since $\mathtt{T}_5(f_5) = \mathtt{N}$, f_5 is NQL. Hence this DAE is NQL.

6 Conclusion

We presented a simple method for quasilinearity analysis when solving a DAE by stages determined from Pryce's SA. Our method is implemented in the DAESA tool [3] for SA of DAEs and the DAETS solver [2]. In DAESA, we also construct a block-triangular form (BTF) of the DAE, and with this analysis, we determine the smallest number of variables and their derivatives that need initial values for a consistent initialization [3]. In DAETS, this method is used to select the appropriate solver when solving up to stage zero.

When applied block-wise to a BTF, our method considers variables that occur in positions outside diagonal blocks as constants. As a result, we need to set the corresponding off-diagonal entries in Θ to $+\infty$ and in $\mathtt{T}$ to $\mathtt{C}$. The propagation rules do not change.

The proof of correctness of our algorithm and a detailed description of how it works in the case of BTFs will be presented in a future work.

References

1. Nedialkov, N.S., Pryce, J.D.: Solving differential-algebraic equations by Taylor series (II): computing the system Jacobian. BIT. **47**(1), 121–135 (2007)
2. Nedialkov, N.S., Pryce, J.D.: Solving differential-algebraic equations by Taylor series (III): the DAETS code. JNAIAM **3**(1–2), 61–80 (2008)
3. Nedialkov, N.S., Pryce, J.D., Tan, G.: DAESA—a MATLAB tool for structural analysis of DAEs: software. ACM Trans. Math. Softw. (2013). Accepted for publication
4. Pryce, J.D.: A simple structural analysis method for DAEs. BIT **41**(2), 364–394 (2001)
5. Pryce, J.D., Nedialkov, N.S., Tan, G.: DAESA—a MATLAB tool for structural analysis of DAEs: Theory. To appear in ACM Transactions on Mathematical Software (2014)

Nondeterministic Fuzzy Operators

Fairouz Tchier

Abstract We consider that nondeterministic programs behave as badly as they can and loop forever whenever they have the possibility to do so. We deal with a relational algebra model to define a nondeterministic refinement fuzzy ordering (*nondeterministic fuzzy inclusion*) and also the associated fuzzy operations which are fuzzy nondeterministic join ($\sqcup_{fuz}$), fuzzy nondeterministic meet ($\sqcap_{fuz}$), and fuzzy nondeterministic composition ($\Box_{fuz}$). We also give some properties of these operations and illustrate them with simple examples.

1 Fuzzy Relations

Fuzzy relations are fuzzy subsets of $A \times B$, that is, mapping from $A \rightarrow B$. They have been studied by a number of authors, in particular by Zadeh [12, 13], Kaufmann [6], and Rosenfeld [10]. Applications of fuzzy relations are widespread and important.

Definition 1 Let $A, B \in U$ be universal sets, a fuzzy relation $\tilde{R}$ on $A \times B$ is defined by:

$\tilde{R} = \{((x, y), \mu_{\tilde{R}}(x, y) \mid (x, y) \in A \times B, \mu_{\tilde{R}}(x, y) \in [0, 1]\}$.

Let $\tilde{R}$ and $\tilde{S}$ be two fuzzy relations on $A \times B$. Then, the following operations are defined:

- *Union*: $\mu_{\tilde{R}\cup\tilde{S}}(x, y) = \mu_{\tilde{R}}(x, y) \vee \mu_{\tilde{S}}(x, y)$
- *Intersection*: $\mu_{\tilde{R}\cap\tilde{S}}(x, y) = \mu_{\tilde{R}}(x, y) \wedge \mu_{\tilde{S}}(x, y)$
- *Max-min composition*:
  $\tilde{R} \circ \tilde{S} \ = \ \{[(x, z), \vee_y\{\mu_{\tilde{R}}(x, y) \wedge \mu_{\tilde{S}}(y, z)\}]\}$.

F. Tchier (✉)
Mathematics Department, P.O. Box 22452, Riyadh 11495, Saudi Arabia
e-mail: ftchier@ksu.edu.sa

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_65

Definition 2 Let $\tilde{R}$ be a fuzzy relation on $A \times A$. Most of these notions are taken from [6, 13]

- $\tilde{R}$ is *reflexive* iff $\mu_{\tilde{R}}(x,x) = 1 \; \forall x \in A$
- $\tilde{R}$ is *transitive* iff $\mu_{\tilde{R}}(x,z) \geq \mu_{\tilde{R}}(x,y) \wedge \mu_{\tilde{R}}(y,z), \; \forall x,y,z \in A$
- $\tilde{R}$ is *symmetric* iff $\tilde{R}(x,y) = \tilde{R}(y,x)$
- $\tilde{R}$ is *antisymmetric* iff for $x \neq y$ either $\mu_{\tilde{R}}(x,y) \neq \mu_{\tilde{R}}(y,x)$ or $\mu_{\tilde{R}}(x,y) = \mu_{\tilde{R}}(y,x) = 0\,, \forall x,y \in A$
- $\tilde{R}$ is *equivalence* iff $\tilde{R}$ is reflexive, transitive, and symmetric
- $\tilde{R}$ is *order* iff $\tilde{R}$ is reflexive, transitive, and antisymmtric

The following properties have been proved to hold for fuzzy relations (see [7, 8]);

Theorem 1
Let $\tilde{R}$, $\tilde{S}$ and $\tilde{T}$ be fuzzy relations. Then:

(a) $\tilde{R}(\tilde{S}\tilde{T}) = (\tilde{R}\tilde{S})\tilde{T}$
(b) $\tilde{R}(\tilde{S} \cup \tilde{T}) = (\tilde{R}\tilde{S}) \cup (\tilde{R}\tilde{T})$
(c) $\tilde{R}(\tilde{S} \cap \tilde{T}) \subseteq (\tilde{R}\tilde{S}) \cap (\tilde{R}\tilde{T})$
(d) $\tilde{S} \subseteq \tilde{T} \Longrightarrow \tilde{R}\tilde{S} \subseteq \tilde{R}\tilde{T}$
(e) $\tilde{S} \subseteq \tilde{T} \Longrightarrow \tilde{S}\tilde{R} \subseteq \tilde{T}\tilde{R}$
(f) $\tilde{R}\tilde{I} = \tilde{I}\tilde{R} = \tilde{R}$ *for all fuzzy relation* $\tilde{R}$
(g) $(\tilde{R}\tilde{S})\smile = \tilde{S}\smile \tilde{R}\smile$
(h) $\tilde{R}\smile\smile = \tilde{R}$
(i) $(\tilde{R} \cup \tilde{S})\smile = \tilde{R}\smile \cup \tilde{S}\smile$
(j) $(\tilde{R} \cap \tilde{S})\smile = \tilde{R}\smile \cap \tilde{S}\smile$
(k) $\tilde{R} \subseteq \tilde{S} \Longrightarrow \tilde{R}\smile \subseteq \tilde{S}\smile$

2 A Nondeterministic Order Refinement

We will give the definition of our ordering.

Definition 3 A relation Q *refines* a relation R [9], denoted by $Q \sqsubseteq R$, iff $RL \subseteq QL$ and $Q \cap RL \subseteq R$, or equivalently, iff $Q \cup \overline{QL} \subseteq R \cup \overline{RL}$ and $\overline{QL} \subseteq \overline{RL}$.

Theorem 2 *The relation $\sqsubseteq$ is a partial order.*

In the following, we will present nondeterministic operators and also some of their properties. For more details see [4, 11]. To clarify the ideas, take two relations Q and R:

- Their supremum is: $Q \sqcup R = (Q \cup R) \cap QL \cap RL$, and satisfies $(Q \sqcup R)L = QL \cap RL$. Then, $Q \sqcup R$ is exactly the relational expression of the *nondeterministic union* as defined by [1, 2] (which explains the word *nondeterministic* of $\sqcup$-semilattice $(\mathcal{B}_R, \sqsubseteq)$).

- Their infimum, if it exists, is $Q \sqcap R = (Q \cup \overline{QL}) \cap (R \cup \overline{RL}) \cap (QL \cup RL)$
$$= Q \cap R \cup Q \cap \overline{RL} \cup R \cap \overline{QL},$$
and it satisfies $(Q \sqcap R)L = QL \cup RL$. The operator $\sqcap$ is called *nondeterministic intersection*. For $Q \sqcap R$ to exist, we have to verify $L \subseteq ((Q \cup \overline{QL}) \cap (R \cup \overline{RL}))L$. This condition is equivalent to $QL \cap RL \subseteq (Q \cap R)L$, which can be interpreted as follows: the existence condition simply means on the intersection of their domains, Q and R have to agree for at least one value.

In what follows, we will give the definition of nondeterministic composition [1–3].

- The binary operator $\triangleright$, called *relative implication*, is defined as follows:
$$Q \triangleright R \overset{\text{def}}{=} \overline{Q\overline{R}}.$$
- The *nondeterministic composition* of relations Q and R is
$$Q \,\Box\, R = QR \cap Q \triangleright RL.$$

The nondeterministic operators $\sqcap$, $\sqcup$, and $\Box$ have the same properties as $\cap$, $\cup$, and (;), but the nondeterministic intersections have to be defined. Let us give some of them.

Theorem 3 *Let P, Q, and R be relations. Then,*

(a) $P \sqcap (Q \sqcup R) = (P \sqcap Q) \sqcup (P \sqcap R)$
(b) $P \sqcup (Q \sqcap R) = (P \sqcup Q) \sqcap (P \sqcup R)$
(c) $R \,\Box\, I = I \,\Box\, R = R$
(d) $Q \sqsubseteq R \Rightarrow P \,\Box\, Q \sqsubseteq P \,\Box\, R$
(e) $P \sqsubseteq Q \Rightarrow P \,\Box\, R \sqsubseteq Q \,\Box\, R$
(f) $P \,\Box\, (Q \sqcup R) = P \,\Box\, Q \sqcup P \,\Box\, R$
(g) $(P \sqcup Q) \,\Box\, R = P \,\Box\, R \sqcup Q \,\Box\, R$
(h) $P \,\Box\, (Q \sqcap R) \sqsubseteq P \,\Box\, Q \sqcap P \,\Box\, R$
(i) $P \,\Box\, (Q \,\Box\, R) = (P \,\Box\, Q) \,\Box\, R$
(j) $(P \sqcap Q) \,\Box\, R \sqsubseteq P \,\Box\, R \sqcap Q \,\Box\, R$
(k) *Q deterministic* $\Rightarrow Q \,\Box\, R = QR$
(l) *P deterministic* $\Rightarrow P \,\Box\, (Q \sqcap R) = PQ \sqcap PR$
(m) *R total* $\Rightarrow Q \,\Box\, R = QR$
(n) $PL \cap QL = \emptyset \Rightarrow (P \cup Q) \,\Box\, R = P \,\Box\, R \cup Q \,\Box\, R$
(o) $PL \cap QL = \emptyset \Rightarrow P \sqcap Q = P \cup Q$

3 A Nondeterministic Fuzzy Order Refinement

We will give the definition of domain of fuzzy relations $\tilde{R}$.

Definition 4 The domain of $\tilde{R}$ is supremum of value in first row of the matrix, and the image of $\tilde{R}$ is supremum of value in first column of the matrix. Formally, $\mathrm{dom}(\tilde{R})d\ ef{=}sup_{y\in B}\{((x, y), \mu_{\tilde{R}}(x, y)) \mid \forall x \in A\}$, $\mathrm{img}(\tilde{R})d\ ef{=}sup_{x\in A}$

$\{((x, y), \mu_{\tilde{R}}(x, y)) \mid \forall y \in B\}$. • The vectors $\tilde{R}\tilde{L}$ and $\tilde{R} \smile \tilde{L}$ are particular vectors characterizing, respectively, the domain and codomain of $\tilde{R}$.

Now, we will give the definition of fuzzy ordering.

Definition 5 We say that a fuzzy relation $\tilde{Q}$ *fuzzy refines* a fuzzy relation $\tilde{R}$, denoted by $\tilde{Q} \sqsubseteq_{fuz} \tilde{R}$, iff $\tilde{R}\tilde{L} \subseteq \tilde{Q}\tilde{L}$ and $\tilde{Q} \cap \tilde{R}\tilde{L} \subseteq \tilde{R}$, i.e, $(\vee_{y \in B}\{\mu_{\tilde{R}}(x, y)\} \leq \vee_{y \in B}\{\mu_{\tilde{Q}}(x, y)\})$ and $(\mu_{\tilde{Q}}(x, y) \wedge (\vee_{y \in B}\{\mu_{\tilde{R}}(x, y)\}) \leq \mu_{\tilde{R}}(x, y))$.

In other words, $\tilde{Q}$ refines $\tilde{R}$, if and only if, the prerestriction of $\tilde{Q}$ to the domain of $\tilde{R}$ is included in $\tilde{R}$. This means that $\tilde{Q}$ must not produce results not allowed by $\tilde{R}$ for those states that are in the domain of $\tilde{R}$.

Theorem 4 *The relation $\sqsubseteq$ is a partial order.*

We will present fuzzy nondeterministic operators and also some of their properties. To clarify the ideas, take two relations $\tilde{Q}$ and $\tilde{R}$:

- Their supremum is $\tilde{Q} \sqcup_{fuz} \tilde{R} = (\tilde{Q} \vee \tilde{R}) \wedge \tilde{Q}\tilde{L} \wedge \tilde{R}\tilde{L}$,
  $\Longleftrightarrow \mu_{(\tilde{Q} \sqcup_{fuz} \tilde{R})}(x, y) = min\{max\{\mu_{\tilde{Q}}(x, y), \mu_{\tilde{R}}(x, y)\}$,
  $max_y(\mu_{\tilde{Q}}(x, y)), max_y(\mu_{\tilde{R}}(x, y))\}$ and satisfies $(\tilde{Q} \sqcup_{fuz} \tilde{R})\tilde{L} = \tilde{Q}\tilde{L} \cap \tilde{R}\tilde{L}$.
  Then, $\tilde{Q} \sqcup_{fuz} \tilde{R}$ is exactly the relational expression of the *fuzzy nondeterministic union*.
- Their infimum, if it exists, is $\tilde{Q} \sqcap_{fuz} \tilde{R} = (\tilde{Q} \wedge \tilde{R}) \vee (\tilde{Q} \wedge 1 - \tilde{R}\tilde{L}) \vee (\tilde{R} \wedge 1 - \tilde{Q}\tilde{L})$
  $\Longleftrightarrow \mu_{(\tilde{Q} \sqcap_{fuz} \tilde{R})}(x, y) = max\{min\{\mu_{\tilde{Q}}(x, y), \mu_{\tilde{R}}(x, y)\}$,
  $min\{\mu_{\tilde{Q}}(x, y), 1 - max_y(\mu_{\tilde{R}}(x, y))\}, min\{\mu_{\tilde{R}}(x, y)$,
  $1 - max_y(\mu_{\tilde{Q}}(x, y))\}\}$ and it satisfies $(\tilde{Q} \sqcap_{fuz} \tilde{R})\tilde{L} = \tilde{Q}\tilde{L} \cup \tilde{R}\tilde{L}$. The operator $\sqcap_{fuz}$ is called *fuzzy nondeterministic intersection*. For $\tilde{Q} \sqcap_{fuz} \tilde{R}$ to exist, we have to verify $\tilde{L} \subseteq (\tilde{Q} \cup \overline{\tilde{Q}}\tilde{L} \cap \tilde{R} \cup \overline{\tilde{R}}\tilde{L})$. This condition is equivalent to $\tilde{Q}\tilde{L} \cap \tilde{R}\tilde{L} \subseteq (\tilde{Q} \cap \tilde{R})\tilde{L}$, which can be interpreted as follows: the existence condition simply means that on the intersection of their domains, $\tilde{Q}$ and $\tilde{R}$ have to agree for at least one value.

In what follows, we will give the definition of the fuzzy nondeterministic composition.

Definition 6 The *fuzzy nondeterministic composition* of relations $\tilde{Q}$ and $\tilde{R}$ is:

$$\tilde{Q} \,\square_{fuz} \tilde{R} = \tilde{Q}\tilde{R} \wedge 1 - \tilde{Q}\overline{\tilde{R}\tilde{L}}$$

$\Longleftrightarrow$

$$\mu_{(\tilde{Q} \,\square_{fuz} \tilde{R})}(x, y) =$$

$min[max_y\{min\{\mu_{\tilde{Q}}(x, y), \mu_{\tilde{R}}(y, z)\}\}, 1 - max_y\{min\{\mu_{\tilde{Q}}(x, y), 1 - max_y(\mu_{\tilde{R}}(x, y))$
$\}\}]$.

The fuzzy nondeterministic operators $\sqcap_{fuz}, \sqcup_{fuz}$, and $\square_{fuz}$, have the same properties as $\sqcap, \sqcup$, and $\square$, but the fuzzy nondeterministic intersections have to be defined. Let us give some of them.

Theorem 5 *Let $\tilde{P}$, $\tilde{Q}$ and $\tilde{R}$ be fuzzy relations. Then,*

- $\tilde{P} \sqcap_{fuz} (\tilde{Q} \sqcup_{fuz} \tilde{R}) = (\tilde{P} \sqcap_{fuz} \tilde{Q}) \sqcup_{fuz} (\tilde{P} \sqcap_{fuz} \tilde{R})$
- $\tilde{P} \sqcup_{fuz} (\tilde{Q} \sqcap_{fuz} \tilde{R}) = (\tilde{P} \sqcup_{fuz} \tilde{Q}) \sqcap_{fuz} (\tilde{P} \sqcup_{fuz} \tilde{R})$
- $\tilde{R} \square_{fuz} \tilde{I} = \tilde{I} \square_{fuz} \tilde{R} = \tilde{R}$
- $\tilde{Q} \sqsubseteq_{fuz} \tilde{R} \Rightarrow \tilde{P} \square_{fuz} \tilde{Q} \sqsubseteq_{fuz} \tilde{P} \square_{fuz} \tilde{R}$
- $\tilde{P} \sqsubseteq_{fuz} \tilde{Q} \Rightarrow \tilde{P} \square_{fuz} \tilde{R} \sqsubseteq_{fuz} \tilde{Q} \square_{fuz} \tilde{R}$
- $\tilde{P} \square_{fuz} (\tilde{Q} \sqcup_{fuz} \tilde{R}) = \tilde{P} \square_{fuz} \tilde{Q} \sqcup_{fuz} \tilde{P} \square_{fuz} \tilde{R}$
- $(\tilde{P} \sqcup_{fuz} \tilde{Q}) \square_{fuz} \tilde{R} = \tilde{P} \square_{fuz} \tilde{R} \sqcup_{fuz} \tilde{Q} \square_{fuz} \tilde{R}$
- $\tilde{P} \square_{fuz} (\tilde{Q} \sqcap_{fuz} \tilde{R}) \sqsubseteq_{fuz} \tilde{P} \square_{fuz} \tilde{Q} \sqcap_{fuz} \tilde{P} \square_{fuz} \tilde{R}$
- $\tilde{P} \square_{fuz} (\tilde{Q} \square_{fuz} \tilde{R}) = (\tilde{P} \square_{fuz} \tilde{Q}) \square_{fuz} \tilde{R}$
- $(\tilde{P} \sqcap_{fuz} \tilde{Q}) \square_{fuz} \tilde{R} \sqsubseteq_{fuz} \tilde{P} \square_{fuz} \tilde{R} \sqcap_{fuz} \tilde{Q} \square_{fuz} \tilde{R}$

Proposition 1

- $\tilde{Q}$ *deterministic* $\Rightarrow \tilde{Q} \square_{fuz} \tilde{R} = \tilde{Q}\tilde{R}$
- $\tilde{P}$ *deterministic* $\Rightarrow \tilde{P} \square_{fuz} (\tilde{Q} \sqcap_{fuz} \tilde{R}) = \tilde{P}\tilde{Q} \sqcap_{fuz} \tilde{P}\tilde{R}$
- $\tilde{R}$ *total* $\Rightarrow \tilde{Q} \square_{fuz} \tilde{R} = \tilde{Q}\tilde{R}$
- $\tilde{P}\tilde{L} \sqcap_{fuz} \tilde{Q}\tilde{L} = \emptyset \Rightarrow (\tilde{P} \sqcup_{fuz} \tilde{Q}) \square_{fuz} \tilde{R} = \tilde{P} \square_{fuz} \tilde{R} \cup \tilde{Q} \square_{fuz} \tilde{R}$
- $\tilde{P}\tilde{L} \sqcap_{fuz} \tilde{Q}\tilde{L} = \emptyset \Rightarrow \tilde{P} \sqcap_{fuz} \tilde{Q} = \tilde{P} \cup \tilde{Q}$

There are many properties achieved for relations, but not fulfilled for fuzzy relations. For instance, if $\tilde{Q}$ and $\tilde{R}$ are fuzzy relations, then:

(a) $\overline{\tilde{Q} \sqcup_{fuz} \tilde{R}} \neq \overline{\tilde{Q}} \sqcap_{fuz} \overline{\tilde{R}}$
(b) $\overline{\tilde{Q} \sqcap_{fuz} \tilde{R}} \neq \overline{\tilde{Q}} \sqcup_{fuz} \overline{\tilde{R}}$
(c) $(\tilde{Q} \sqcap_{fuz} \tilde{R}) \sqcup_{fuz} \tilde{R} \neq \tilde{Q} \sqcup_{fuz} \overline{\tilde{R}}$
(d) $\tilde{Q} \sqsubseteq_{fuz} \tilde{R} \not\Rightarrow \overline{\tilde{R}} \not\sqsubseteq_{fuz} \overline{\tilde{Q}}$

Example Let $\tilde{Q} = \begin{pmatrix} 0.1 & 0 \\ 1 & 0.2 \end{pmatrix}$, $\tilde{R} = \begin{pmatrix} 0.2 & 0.3 \\ 0.4 & 0.8 \end{pmatrix}$

- $\overline{\tilde{Q} \sqcup_{fuz} \tilde{R}} = \begin{pmatrix} 0.9 & 0.9 \\ 0.2 & 0.2 \end{pmatrix}$ but $\overline{\tilde{Q}} \sqcap_{fuz} \overline{\tilde{R}} = \begin{pmatrix} 0.8 & 0.7 \\ 0.2 & 0.6 \end{pmatrix}$
- $\overline{\tilde{Q} \sqcap_{fuz} \tilde{R}} = \begin{pmatrix} 0.8 & 0.7 \\ 0.4 & 0.8 \end{pmatrix}$ but $\overline{\tilde{Q}} \sqcup_{fuz} \overline{\tilde{R}} = \begin{pmatrix} 0.8 & 0.8 \\ 0.4 & 0.4 \end{pmatrix}$
- $(\tilde{Q} \sqcap_{fuz} \tilde{R}) \sqcup_{fuz} \tilde{R} = \begin{pmatrix} 0.3 & 0.3 \\ 0.4 & 0.2 \end{pmatrix}$ but $\tilde{Q} \sqcup_{fuz} \overline{\tilde{R}} = \begin{pmatrix} 0.1 & 0.1 \\ 0.4 & 0.2 \end{pmatrix}$

- Let $\tilde{Q} = \begin{pmatrix} 0.1 & 0 \\ 1 & 0.2 \end{pmatrix}$, $\tilde{R} = \begin{pmatrix} 0.1 & 0.1 \\ 0.4 & 0.4 \end{pmatrix}$ then $\tilde{Q} \sqsubseteq_{fuz} \tilde{R}$ but $\overline{\tilde{R}} \not\sqsubseteq_{fuz} \overline{\tilde{Q}}$ because
$\overline{\tilde{Q}}\tilde{L} = \begin{pmatrix} 1 \\ 0.8 \end{pmatrix} \not\subseteq \begin{pmatrix} 0.9 \\ 0.6 \end{pmatrix} = \overline{\tilde{R}}\tilde{L}.$

References

1. Berghammer, R.: Relational Specification of Data Types and Programs. Technical report 9109, Fakultät für Informatik, Universität der Bundeswehr München, Germany, Sept. (1991)
2. Berghammer, R., Schmidt, G.: Relational specifications. In: Rauszer, C. (ed.) Proc. XXXVIII Stefan Banach Seminar on Algebraic Methods in Logic and Their Computer Science Applications, Vol. 28, pp. 167–190. Banach Center Publications, Institute of Mathematics, Polish Academy of Sciences, Warsaw (1993)
3. Berghammer, R., Zierer, H.: Relational algebraic semantics of deterministic and nondeterministic programs. Theor. Comput. Sci. **43**, 123–147 (1986)
4. Desharnais, J., Tchier, F.: Demonic relational semantics of sequential programs. Rapport de recherche DIUL-RR-9406, Departement d'Informatique, Universite Laval, Quebec, Canada, decembre (1995)
5. Desharnais, J., Belkhiter, N., Ben Mohamed Sghaier, S., Tchier, F., Jaoua, A., Mili, A., Zaguia, N.: Embedding a demonic semilattice in a relation algebra. Theor. Comput. Sci. **149**(2):333–360 (1995)
6. Kaufmann, A.: Introduction to the Theory of Fuzzy Subsets, Vol. 1. Elements of Basic Theory, Masson, Paris (1973); Vol. 2. Applications to Linguistics, Logic and Semantics, Masson, Paris (1975); Vol. 3. Applications to Classification and Pattern Recognition, Automata and Systems, and Choice of Criteria, Masson, Paris (1975); also English translation of Vol. 1, Academic, New York (1975)
7. Kaufmann, A.: Introduction à la Théorie des Sous-Ensembles Flous, Vol. IV. Masson, Paris (1977)
8. Kawahara, Y., Furusawa, H.: An algebraic formatisation of fuzzy relations. Presented at Second International Seminar on Relational Methods in Computer Science, Rio, Brazil, July (1995)
9. Mili, A., Desharnais, J., Mili, F.: Relational heuristics for the design of deterministic programs. Acta Inf. **24**(3), 239–276 (1987)
10. Rosenfeld, A.: Fuzzy graphs. In: Zadeh, L. A., Fu, K. S., Shimura, M. (eds.) Fuzzy Sets and Their Applications, pp. 77–95. Academic Press, New York, NY, USA (1975)
11. Tchier, F.: Sémantiques relationnelles démoniaques et vérification de boucles non déterministes. Theses of doctorat, Département de Mathématiques et de statistique, Université Laval, Canada (1996)
12. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
13. Zadeh, L.A.: Similarity relations and fuzzy orderings. Inf. Sci. **3**, 177–206 (1971)

The Ideal Free Distribution and Evolutionary Stability in Habitat Selection Games with Linear Fitness and Allee Effect

Ross Cressman and Tan Tran

Abstract Fretwell and Lucas [3] introduced the Ideal Free Distribution (IFD) to predict how birds establish themselves among habitats. It has been shown that the IFD is an evolutionarily stable strategy (ESS) of the habitat selection game when fitness is a decreasing function of patch density. We develop a formula for the IFD when there are an arbitrary number of habitats, and fitness functions are linearly decreasing in the population size (i.e., density) in each habitat. We also explore the IFD when fitness functions increase with population size until some maximum threshold is reached (Allee Effect) and examine whether an IFD still is an ESS in this case.

1 Habitat Selection Games with Linear Fitness

Fretwell and Lucas [3] initially consider habitats with decreasing suitability (i.e., fitness) as the density in a given habitat increases. They prove that, for a fixed total population size, a unique Ideal Free Distribution (IFD) exists but do not provide a formula for the IFD. In this section, we develop a formula when the suitability functions are linear and decreasing. Suppose fitness f_i in the ith habitat is given by

$$f_i(x_i) = b_i - a_i x_i. \tag{1}$$

The notation in this equation is:

- b_i is the basic suitability for the ith habitat
- a_i is the factor for linear decrease for the ith habitat
- x_i is the density for the ith habitat (cannot be negative)

R. Cressman (✉) · T. Tran
Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada
e-mail: rcressman@wlu.ca

T. Tran
e-mail: tatran@wlu.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_66

Let N denote the total population density. Then, when there are H habitats in total, $N = x_1 + x_2 + x_3 + ... + x_H$ and $p_i = \frac{x_i}{N}$ is the proportion of the population in the ith habitat.

Definition 1 (Ideal Free Distribution) When total density is fixed at N, the IFD is a distribution for which the fitness in all occupied patches is the same and at least as high as the fitness in any unoccupied patch. That is, the IFD satisfies

$$\begin{aligned} f_i(x_i) &= f_j(x_j) && \text{if } x_i, x_j > 0 \text{ and} \\ f_i(x_i) &\geq f_j(x_j) && \text{if } x_i > 0 \text{ and } x_j = 0. \end{aligned}$$

This definition is equivalent to defining the IFD as a Nash equilibrium (NE) of the H-strategy habitat selection game. This is a game between an individual and the population average strategy [1, 7]. In fact, it has been shown that the IFD is unique for fixed population density N [3] and that the IFD is also an ESS [2] when fitness is a decreasing function of density in each habitat.

When fitness is linearly decreasing as in (1), the habitat selection game is represented by the $H \times H$ payoff matrix A given by

$$A = \begin{bmatrix} b_1 - Na_1 & b_1 & b_1 & ... & b_1 \\ b_2 & b_1 - Na_2 & b_2 & ... & b_2 \\ ... & ... & ... & ... & ... \\ b_H & b_H & b_H & ... & b_H - Na_H \end{bmatrix}. \tag{2}$$

That is, with e_i the unit vector whose ith component is 1 and all others are 0,

$$e_i \cdot Ap = b_i(p_1 + ... + p_N) - Na_i p_i = b_i - a_i x_i = f_i(x_i). \tag{3}$$

This matrix notation for fitness proves useful in the remainder of the section.

Example 1 (Two Habitats) Let $f_1(x_1) = 2 - 3x_1$ and $f_2(x_2) = 1 - 2x_2$ be the fitness functions in the two habitats. Their graphs are given in Fig. 1a. Here we follow Fretwell and Lucas [3] in that the basic suitability in habitat 1 is larger than that of habitat 2 (i.e., $b_1 > b_2$). Since patch 1 is the better habitat (i.e., $b_1 > b_2$) at low densities, all individuals will go to patch 1 until total population size reaches some threshold level. To find this threshold N^*, we solve when the fitness in patch 1 equals that in an unoccupied patch 2 (i.e., $e_1 \cdot Ap = e_2 \cdot Ap$ where $p = (p_1^*, p_2^*) = (1, 0)$).

Since $A = \begin{bmatrix} 2 - 3N & 2 \\ 1 & 1 - 2N \end{bmatrix}$ in (2), we find that

$$[1 \ \ 0\,] \begin{bmatrix} 2 - 3N & 2 \\ 1 & 1 - 2N \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = [0 \ \ 1] \begin{bmatrix} 2 - 3N & 2 \\ 1 & 1 - 2N \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Leftrightarrow N = \frac{1}{3}.$$

Therefore, above the threshold $N^* = \frac{1}{3}$, the species occupy both patch 1 and patch 2 in order to equalize fitness in both habitats. For $N > N^*$, from $e_1 \cdot Ap^* = e_2 \cdot Ap^*$

and $p_2^* = 1 - p_1^*$, we find that $p_1^* = \frac{1+2N}{5N}$ and $p_2^* = \frac{3N-1}{5N}$. That is, the IFD (p_1^*, p_2^*) is given as follows in terms of N:

$$p_1^* = \begin{cases} 1, & N \le \dfrac{1}{3} \\ \dfrac{1+2N}{5N}, & N > \dfrac{1}{3} \end{cases} \quad \text{and } p_2^* = \begin{cases} 0, & N \le \dfrac{1}{3} \\ \dfrac{3N-1}{5N}, & N > \dfrac{1}{3} \end{cases} \tag{4}$$

These are plotted in Fig. 1b. In this example, it is interesting to note that as N increases (specifically for $N > 2$), more of the population is in patch 2 than in patch 1 (i.e., $p_2^* > p_1^*$) even though patch 1 is better than patch 2 at low densities.

Note that in this example and throughout the article, fractional values appear for N and p_i. An integer number of individuals arise in each patch if population sizes are measured in large units (e.g., in thousands).

The method in Example 1 can be generalized to $H \ge 2$ habitats and arbitrary $b_1 > b_2 > b_3 > \ldots > b_H$. Using (2) and solving $e_1 \cdot Ap = e_2 \cdot Ap = \ldots = e_M \cdot Ap$ for N with $p = (p_1^*, p_2^*, \ldots, p_{M-1}^*, 0, \ldots, 0)$ and $M = 2, 3, \ldots, H$ yields the threshold N_M^* given in (5) for the species to begin to occupy habitat M.

$$N_M^* = \sum_{i=1}^{M} \frac{(b_i - b_M)}{a_i} \tag{5}$$

Moreover, for $N_M^* < N < N_{M+1}^*$, the IFD is $(p_1^*, p_2^*, p_3^*, ..., p_H^*)$ where

$$p_k^* = \begin{cases} \dfrac{\frac{1}{a_k} \cdot \left[\sum\limits_{i=1}^{M} \frac{1}{a_i}(b_k - b_i) \right] \cdot \left[\prod\limits_{j=1}^{M} a_j \right] + \frac{N}{a_k} \prod\limits_{j=1}^{M} a_j}{N \left[\sum\limits_{i=1}^{M} (\prod\limits_{j=1}^{M} \frac{a_j}{a_i}) \right]}, & 1 \le k \le M \\ 0, & k > M. \end{cases} \tag{6}$$

Note that, if we define $N_1^* = 0$ and $N_{H+1}^* = \infty$, then (6) is the IFD for any fixed population size $N > 0$. Also, in Example 1, $N_2^* = \frac{b_1 - b_2}{a_1} = \frac{1}{3}$ and (6) simplifies to (4). From (6), it is also clear what happens when population size becomes large. Eventually, all habitats are occupied and the proportion in the kth habitat approach $p_k^{*\infty} = 1/(a_k \sum_{i=1}^{H} \frac{1}{a_i})$. Since a_k measures the density (or logistic) effect in habitat k, we see that the proportion of the population in habitat k decreases as this density effect becomes more pronounced. In particular, there will eventually be more individuals in the patch with the smaller a_i as in Example 1 where $a_2 < a_1$ (See Fig. 1b).

2 Allee Effect—Two Habitat System

Fretwell and Lucas [3] also briefly considered what happens when there are two patches and an Allee Effect (i.e., when patch fitness increases to some threshold K for lower patch density before it decreases as patch density gets higher). In this

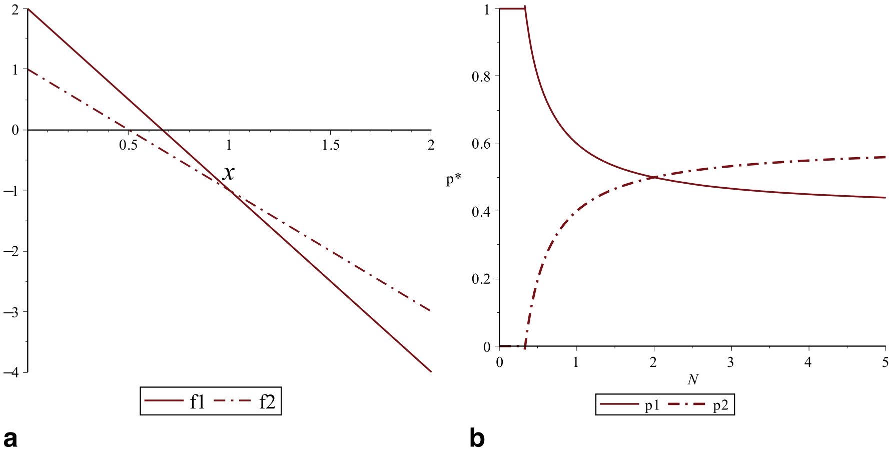


Fig. 1 Example 1. Panel (**a**) plots the suitability functions. Panel (**b**) is the Ideal Free Distribution as a function of N

section, we examine this question when patch fitness is a piecewise linear function, a continuous function that increases to some maximum value and decreases thereafter.

The general piecewise linear two-habitat model with Allee Effect has fitness functions of the form

$$f_1(x_1) = \begin{cases} m_{11}x_1 + b_1 & 0 \le x_1 \le \frac{K_1-b_1}{m_{11}} \\ K_1 + m_{12}(\frac{K1-b_1-x_1m_{11}}{m_{11}}) & x_1 \ge \frac{K_1-b_1}{m_{11}} \end{cases}$$

$$f_2(x_2) = \begin{cases} m_{21}x_2 + b_2 & 0 \le x_2 \le \frac{K_2-b_2}{m_{21}} \\ K_2 + m_{22}(\frac{K2-b_2-x_2m_{21}}{m_{21}}) & x_2 \ge \frac{K_2-b_2}{m_{21}}, \end{cases}$$

where the threshold $K_i > b_i$ in each patch and $b_1 > b_2$ so that patch 1 is still the better patch at low density (See Fig. 2a). As stated in Sect. 1, a NE of this two-patch habitat selection game at fixed total population size N is given by Definition 1. These can occur when both patches are occupied (i.e., $f_1(x_1) = f_2(x_2)$ where $x_1+x_2 = N$) or where the species is all in one patch (i.e., either $f_1(N) \ge f_2(0)$ or $f_2(N) \ge f_1(0)$). These criteria are applied to the following example.

Example 2 Letting $b_1 = 2, b_2 = 1, K_1 = 5, K_2 = 5, m_{11} = 1, m_{12} = 1, m_{22} = 2,$ $m_{21} = 2$ yields the fitness functions

$$f_1(x_1) = \begin{cases} x_1 + 2 & 0 \le x_1 \le 3 \\ 8 - x_1 & x_1 \ge 3 \end{cases} \text{ and } f_2(x_2) = \begin{cases} 2x_2 + 1 & 0 \le x_2 \le 2 \\ 9 - 2x_2 & x_2 \ge 2 \end{cases}.$$

A tedious calculation produces the following NE shown in Fig. 2b.

If $0 \le N \le 7$, then $p^* = (1,0)$ is a NE. If $\frac{1}{2} \le N \le \frac{7}{2}$, then $p^* = (0,1)$ is a NE.

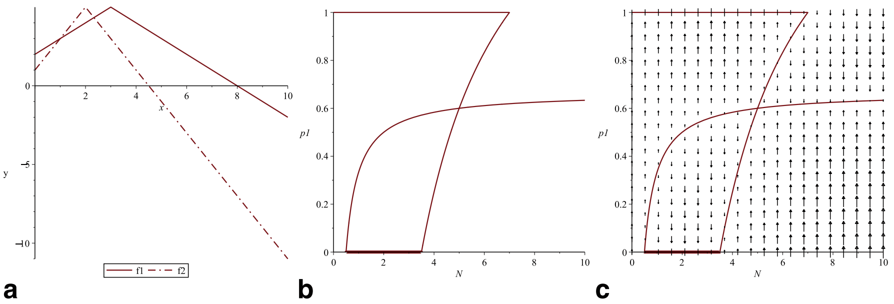


Fig. 2 Suitability functions (**a**), Nash Equilibrium (**b**) and fieldplot for the replicator equation (**c**) for Example 2 and total population size between 0 and 10

If $\frac{7}{2} \le N \le 7$, then $p^* = (\frac{2N-7}{N}, \frac{7-N}{N})$ is a NE. If $N \ge \frac{1}{2}$, then $p^* = (\frac{2N-1}{3N}, \frac{N+1}{3N})$ is a NE.

Similar to Sect. 1, only the better patch is occupied at low population size ($0 \le N < \frac{1}{2}$) and there is a unique NE with both patches occupied when this size is large enough ($N > 7$). These properties remain true for all two habitat models with or without the Allee Effect. However, at intermediate population sizes, models with the Allee Effect display two new phenomena. First, a NE can occur where only the worse patch is occupied ($\frac{1}{2} \le N \le \frac{7}{2}$) and the NE is not always unique for a fixed population size ($\frac{1}{2} \le N \le 7$). In this example, there are up to three NE at a fixed N.

Whenever an evolutionary game has multiple NE, the question of their stabilities arises. Typically, only those that are either an ESS or dynamically stable under an evolutionary game dynamics are considered to be selected as the evolutionary outcome [9]. Since the habitat selection game at fixed N of this section is a two-strategy population game, a strategy p^* is an ESS if and only if it is dynamically stable under the replicator equation [4] given by the (one-dimensional) differential Eq. (7) where $0 \le p_1 \le 1$.

$$\dot{p_1} = p_1(1 - p_1)(f_1(Np_1) - f_2(Np_2)) \tag{7}$$

Figure 2c shows the fieldplot (i.e., the direction of the vector field) of (7) at each fixed N for Example 2. For $\frac{1}{2} \le N \le 7$, there are exactly two ESS's given by the top curve $p^* = (1, 0)$ and the piecewise-defined bottom curve in Fig. 2b, c. Otherwise, there is a unique ESS with all individuals in the better patch ($p^* = (1, 0)$ for $0 < N < \frac{1}{2}$) or both patches are occupied with $p^* = (\frac{2N-1}{3N}, \frac{N+1}{3N})$ for $N > 7$. In fact, it has been shown [5] that up to three ESSs can coexist in a two-patch model with Allee Effect if parameters are chosen properly (e.g., $b_1 = -0.375, b_2 = -1.05, K_1 = \frac{259}{136}, K_2 = \frac{129}{80}, m_{11} = 0.25, m_{12} = \frac{7}{40}, m_{21} = 0.3, m_{22} = 0.1$).

3 Discussion

It is well-known [3] that the IFD $p^* = (p_1^*, \ldots, p_H^*)$ is unique when population density N is fixed and fitness is a decreasing function of density in each patch. However, an explicit formula for p^* in terms of N has only been given for two-habitat models (i.e., $H = 2$) and fitness functions that correspond to logistic density dependence in that they are linearly decreasing [6]. Section 1 extends this result to an arbitrary number of habitats with linear fitness by providing the threshold density (5) at which each patch begins to be occupied and describing how the population is distributed (6) among the occupied patches at a given N. These equations show explicitly how basic suitabilities (i.e., the fitness of an unoccupied patch) and differing patch carrying capacities affect the IFD.

According to the original definition of Fretwell and Lucas [3] (see Definition 1 in Sect. 1), an IFD corresponds to a NE of the habitat selection game [2]. As shown in Sect. 2 where fitness is piecewise linear, several such IFD can already emerge at the same population density in two-habitat models. In fact, the patch that is better at low density may be completely unoccupied at such an IFD. On the other hand, it has also been recognized [3, 8] that these NE may be unstable since, if one individual shifts to the other patch when the population is at one of the NE distributions, more will follow since their fitness will increase by doing so.

A similar phenomenon in two-species habitat selection models led Krivan et al. [6] to define an IFD to be a distribution that is stable under an evolutionary dynamics that models the consequences of these perturbations from NE. Applied to our two-habitat single species, this amounts to selecting IFD's as the NEs that are stable under the replicator Eq. (7) (see also [5]). These also correspond to the ESSs of the habitat selection game (Sect. 2). A different approach to the stability issue was taken by Fretwell and Lucas [3] and Morris [8]. They start at low density and allow this to increase by introducing one new individual at a time. The stable distribution is then determined by finding which patch this new individual would choose to occupy and what cascading effect (if any) this has on choices of individuals currently in the population. This method provides some of the dynamically stable NE, but not necessarily all of them. For example, Morris ([8]; Fig. 2) finds a unique stable NE for each population density whereas another ESS emerges by starting the population in the patch that is worse at low density. Extensions to more than two patches that include Allee Effects add further complications to defining and determining the IFDs that are beyond the scope of this chapter.

References

1. Broom, M., Rychtar, J.: Game-Theoretical Models in Biology. CRC Press, Oxfordshire (2013)
2. Cressman, R., Krivan, V.: Migration dynamics for the ideal free distribution. Am. Nat. **168**, 384–397 (2006)
3. Fretwell, S., Lucas, H.: On territorial behavior and other factors influencing habitat distribution in birds. Acta Biotheor. **19**, 16–32 (1969)

4. Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge (2010)
5. Krivan, V.: The Allee-type ideal free distribution. J. Math. Biol. **69**, 1497–1513 (2014)
6. Krivan, V., Cressman, R., Schneider, C.: The ideal free distribution: a review and synthesis of the game-theoretic perspective. Theor. Popul. Biol. **73**, 403–425 (2008)
7. Maynard Smith, J.: Evolution and the Theory of Games. Cambridge University Press, Cambridge (1982)
8. Morris, D.: Measuring the Allee effect: positive density dependence in small mammals. Ecology **83**, 14–20 (2002)
9. Samuelson, L.: Evolutionary Games and Equilibrium Selection. MIT, Cambridge (1997)

An Input–Output Analysis Approach in Waste of Electrical and Electronic Equipments

Ziya Ulukan, Emre Demircioglu and Mujde Erol Genevois

Abstract The disposal of waste of electrical and electronic equipments (WEEE) represents the loss of large amounts of valuable resources, in particular metals and plastics. If these were to be recycled, it would not only divert the waste from disposal by limiting waste flows damage but would also reduce the need to use virgin raw materials. In this study, we focus on waste management and we concentrate on the recycling of mobile phones. Mobile phone components and their requirements in production phases such as energy, labor, and know-how are depicted in a matrix form inspired from the seminal work of Leontief and input–output (I–O) methodology which appears to be appropriate for analyzing waste management problem is presented. Thus, we propose numerically static I–O solutions in order to demonstrate the contribution of recycling of mobile phones into the economy. Particularly, we concentrate on the monetary IO table (MIOT) and environmental IO table (EIOT) with recycling and balance equations. By defining waste outputs as a new vector class, the classical I–O model has been improved.

1 Introduction

Input–output (I–O) analysis is a method of systematically quantifying the mutual interrelationships among the various sectors of a complex economic system. In practical terms, the economic system to which it is applied may be as large as a nation or even the entire world economy, or as small as the economy of a metropolitan area or even a single enterprise [1].

I–O models have some major extensions in literature as a physical IO table (PIOT), monetary IO table (MIOT), waste IO table (WIOT) and environmental IO table

Z. Ulukan (✉) · E. Demircioglu · M. Erol Genevois
Galatasaray University, Ortakoy, Istanbul, Turkey
e-mail: zulukan@gsu.edu.tr

E. Demircioglu
e-mail: edemircioglu@gsu.edu.tr

M. Erol Genevois
e-mail: merol@gsu.edu.tr

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_67

(EIOT). The first one provides a framework in which all physical flows associated with an economy can be recorded, while the second gives an insight into the value of economic transactions between different sectors. The third represents the interdependence between the flow of goods and the flow of wastes and the last emphasizes a suitable tool for estimating the short-term response of emissions and resource usage to changes in production induced by economic growth [2].

2 Methodology

This work is conducted by MIOT. It can be seen that both for metals in a mobile phone and for labor force required in manufacturing process, the general terms must be monetary.

Our model is based on the equations of the static multisector I–O model [1]. Its basic notation and fundamental relationships are given by:

$$X_i = \sum_{j=1}^{n} (m_{ij}) + Y_i, \quad for \quad i = 1, 2, ..., n \tag{1}$$

where X_i denotes the total output of sector i, m_{ij} represents the flow of input from sector i to sector j, and Y_i shows the final demand for sector i's production. Equation (1) guarantees that the total output of any sector are consumed by either itself or other sectors and also used up by the demand in that economic system.

By determination a technical coefficient as $a_{ij} = m_{ij}/X_j$, Eq. (1) can be modified and rewritten as follows:

$$X_i - \sum_{j=1}^{n} (a_{ij}) = Y_i, \quad for \quad i = 1, 2, ..., n \tag{2}$$

The last equation can be transformed in a matrix form as below:

$$(I - A).X = Y \tag{3}$$

and so,

$$X = (I - A)^{-1}.Y \tag{4}$$

where A is the technical coefficient matrix with coefficients a_{ij} that identifies the percentage of the total inputs of a sector required to be purchased from another sector and I is the identity matrix of size n. $(I - A)^{-1}$ is a square matrix commonly designated as Leontief inverse, if it exists.

The main difference of the modified version is that the inputs of a manufacturing process do not necessarily have to be equal to the outputs. Namely, some outputs are defined as waste, assuming that there is no way to reuse them. The second difference is that as we focus on a manufacturing process, the structure of the I–O analysis needs to be modified and the factors in a matrix must be expressed in monetary units [3].

Table 1 Monetary input –output table of a manufacturing process

		Outputs											
		S_1^*	S_2^*	...	S_j^*	...	S_n^*	Y_1	Y_2	...	Y_j	...	Y_m
Inputs	S_1	c_{11}	c_{12}	...	c_{1j}	...	c_{1n}	w_{11}	w_{12}	...	w_{1j}	...	w_{1m}
	S_2	c_{21}	c_{22}	...	c_{2j}	...	c_{2n}	w_{21}	w_{22}	...	w_{2j}	...	w_{2m}
	...	...	...	...	...	...	...	...	...	...	...	...	...
	S_i	c_{i1}	c_{i2}	...	c_{ij}	...	c_{in}	w_{i1}	w_{i2}	...	w_{ij}	...	w_{im}
	...	...	...	...	...	...	...	...	...	...	...	...	...
	S_n	c_{n1}	c_{n2}	...	c_{nj}	...	c_{nn}	w_{n1}	w_{n2}	...	w_{nj}	...	w_{nm}

Within a manufacturing process based on the mass conservation, there are n inputs, denoted by S_1 through S_n and $n+m$ outputs denoted by S_1^* through S_n^* with Y_1 through Y_m. Among the outputs, S^* are the original substance form obtained by a process, while Y_n are the new substances referred as a waste. This transaction can be depicted as shown in Table 1 where c_{ij} (arise from cost) signifies the amount of money of the substance S_i that is used for substance S_j^*, while w_{ij} (arise from waste cost) signifies the amount of money of waste Y_j that is transformed from substance S_i after the manufacturing process.

Assuming that there is a linear relationship between inputs and outputs, the following equation can be formulated:

$$X_i = \sum_{j=1}^{n}(c_{ij}) + \sum_{j=1}^{m}(w_{ij}), \quad for \quad i = 1, 2, ..., n \tag{5}$$

where x_i stands for an amount of money of the substance i. a_{ij} and b_{ij} can be stated as original outputs and waste outputs, respectively, that are indicated below:

$$a_{ij} = \frac{c_{ij}}{x_j} \quad and \quad b_{ij} = \frac{w_{ij}}{y_j} \quad for \quad j = 1, 2, ..., m \tag{6}$$

where y_j stands for an amount of money of waste j. Note that both x_j and y_j are computed as a total amount of money for substance j (S_1 through S_n) and waste j (Y_1 through Y_n), respectively.

Using the relationships from the last two equations, following equation system can be hold:

$$X_i = \sum_{j=1}^{n}(a_{ij}.x_j) + \sum_{j=1}^{m}(b_{ij}.y_j), \quad for \quad i = 1, 2, ..., n \tag{7}$$

which can be reformulated in matrix form:

$$X = A.X + B.Y \tag{8}$$

Hence,

$$(I - A).X = B.Y \tag{9}$$

where $X = [x_1 \quad x_2 \quad \dots \quad x_n]^T$, $Y = [y_1 \quad y_2 \quad \dots \quad y_m]^T$,

A and B are technical coefficient matrices that characterize material flow in the process. For instance, a_{11} indicates the ratio of the substance S_1 which is used for the same substance as an output S_1^* and if this coefficient is less than 1, this means that any waste as an output will be created. Note that for every i and j, the technical coefficients have to be nonnegative and less than or equal to 1. Since A is a square matrix and assuming that $(I - A)$ is invertible, the inputs X can be calculated in terms of money as long as the waste outputs Y are known a priori, as below:

$$X = (I - A)^{-1}.B.Y \tag{10}$$

An input can remain unchanged in its amount during the process. In other words, in case any substance is entirely consumed, no waste will be generated. Hence, if this situation occurs, then $I - A$ will become singular where its inverse does not exist. For such a case, unchanged input can be removed from I–O table so that invertibility is guaranteed and Eq. (10) can be computed. Moreover, Eq. (10) can be modified as follows:

$$Y = B^{-1}(I - A).X \tag{11}$$

It should be denoted that matrix B has to be square and invertible in Eq. (11). Condition that square matrix B will become singular commonly occurs by adding an output column that no waste is generated. Then, we apply a new formula as:

$$\bar{Y} = (B^T B)^{-1}.B^T.(I - A).X \tag{12}$$

if $B^T B$ is non-singular [3].

3 Case Study for a Mobile Phone's Manufacturing Process

We can now implement our methodology based on an I–O analysis in order to demonstrate numerically the impact of recycling of a typical mobile phone during its manufacturing process with respect to MIOT.

For simplicity, components of a typical mobile phone are basically divided into four groups: valuable metals, other metal components, plastics, paint. We also take into consideration other requirements such as energy, water, tools, and labor.

where x_1 through x_8 represent the unit cost of valuable metals, other metal components, plastics, paint, energy, water, tools, and labor and y_1 through y_8 represent scrap, fuel, waste chemicals, waste paint, CO_2, waste H_2O, steam, and loss of energy/labor for a typical mobile phone manufacturing process.

Based on a data acquired from a mobile phone firm, Table 2 depicts the inputs and outputs from a manufacturing system of a mobile phone. The values are in terms of money for a typical mobile phone that costs totally 100 monetary units.

Table 2 Input–output transaction in manufacturing process in terms of money per 100 monetary units

	(x_1)	(x_2)	(x_3)	(x_4)	(x_5)	(x_6)	(x_7)	(x_8)	(y_1)	(y_2)	(y_3)	(y_4)	(y_5)	(y_6)	(y_7)	(y_8)	**TOTAL**
Valuable metals (x_1)	15.4								2.6		1.1						19.1
Other metal components (x_2)		22.3								1.3		1					24.6
Plastics (x_3)			15.1							2.5	0.9		1.4				19.9
Paint (x_4)				0.3								2.3					2.6
Energy (x_5)											0.1		4.7			4.6	9.4
Water (x_6)														2.4	1.1		3.5
Tools (x_7)									3.3								3.3
Labor (x_8)																17.6	17.6
Total	15.4	22.3	15.1	0.3	0	0	0	0	5.7	2.5	3.1	2.3	5.6	2.9	2.6	22.2	100

Using the data in Table 2 owing to Eq. (6), the coefficients matrices A and B respectively are obtained as:

$$\begin{pmatrix} 0.81 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.91 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.76 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0.36 & 0 & 0.35 & 0 & 0 & 0 & 0 & 0 \\ 0.18 & 0 & 0.32 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.29 & 0 & 0.23 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.03 & 0 & 0.77 & 0 & 0 & 0.21 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0.46 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.79 \end{pmatrix}$$

Thus, the relation between X (inputs–outputs) and Y (outputs) is given by Eq. (10) as :

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{pmatrix} 1.87 & 0 & 1.83 & 0 & 0 & 0 & 0 & 0 \\ 1.93 & 0 & 3.45 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4.15 & 1.2 & 0 & 0.95 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.03 & 0 & 0.77 & 0 & 0 & 0.21 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0.46 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.79 \end{pmatrix} . \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} \tag{13}$$

Considering Eq. (13), the relationships between any output(s) and total input can be revealed. For instance, if it is desirable to reduce CO_2 emissions, the relationship between corresponding inputs can be modified (Table 3).

Table 3 Summary of multilateral comparisons consisting of two or more output with total input

Output1	Output2	Results	Inputs consumed						Percentage Range of Bound	Critical Percentages		
			Input1	**Input2**	**Input3**	**Input4**	**Input5**	**Input6**		**Bound**	**Ave.**	**Bound**
										Extreme	**(x;x)**	**Extreme**
y1	y2	Total input	x1	x2	x3	x7			20–24	0.24	0.22	0.2
y3	y4	Total input	x1	x2	x3	x4	x5		17–62	0.62	0.26	0.17
y5	y6	Total input	x3	x5	x7				58–93	0.93	0.72	0.58
$y5 + y6$	y8	Total input	x3	x5	x6	x8			74–97	0.74	0.84	0.97
$y1 + y2$	y3	Total input	x1	x2	x3	x5	x7		16–21	0.16	0.18	0.21
y6	y7	Total input	x5	x6	x7				100	1	1	1
$y1 + y2$	$y3 + y4$	Total input	x1	x2	x3	x4	x5	x7	22–26	0.26	0.24	0.22

4 Conclusion

The scope of this chapter and the aim of proposed model were to apply I–O analysis in waste management. This model is referred as monetary environmental I–O model for a manufacturing process and it is the first study for conducting a recycling assessment of mobile phones focusing on its manufacturing process, so there is no chance to compare with any solutions in the literature. In other words, integrating the manufacturing process, the monetary I–O analysis and the recycling of a mobile phone are original parts of this work. As a future work, we attempt to analyze the life cycle of a mobile phone by using the dynamic Leontief model.

Acknowledgement This research has been financially supported by Galatasaray University Research Fund (12.402.007).

References

1. Leontief, W.: Input–Output Economics. Oxford University Press, New York (1986)
2. O'Doherty, J., Tol, R.S.J.: An environment input-output model for Ireland. Econ. Soc. Rev. **38**(2):157–190 (2007)
3. Xue, H., Kumar, V., Sutherland, J.W.: Material flows and environmental impacts of manufacturing systems via aggregated input output models. J. Clean. Prod. **15**(3):1349–1358 (2007)

A Free Boundary Approach to Solve the Equilibrium Equations of a Membrane

Giuseppe Viglialoro, Álvaro González and Juan Murcia

Abstract This chapter deals with a mathematical problem related to the equilibrium analysis of a membrane with rigid and cable boundary. The membrane and its boundary are respectively identified with a regular surface and a set of regular curves. The equilibrium is directly expressed by means of an elliptic problem, in terms of the shape of the membrane and its stress tensor; therefore, a free boundary numerical resolution procedure is presented and applied in a particular case.

1 Introduction

This work studies the bi-dimensional and continuous equilibrium of a membrane for the *prestressing phase*; more exactly, a new membrane technology for footbridges is being developed in Spain and a membrane footbridge prototype has been built (see Fig. 1a). The prestressing is introduced and obtained by means of the membrane's boundaries, that are one-dimensional elements defined by spatial curves; herein both rigid and cable boundaries will be considered.

This author acknowledges *Sardinia Regional Government for the financial support (P.O.R. Sardegna, 2007–2013)* and the Project MTM2010-16401 of the *Ministerio de Ciencia e Innovación/FEDER* and the *Consejería de Economía, Innovación y Ciencia* of the *Junta de Andalucía*.

G. Viglialoro (✉)
Department of Mathematics and Computer Science,
University of Cagliari, Cagliari, Italy
e-mail: giuseppe.viglialoro@unica.it

Á. González
Department of Mathematics, University of Cadiz, Cadiz, Spain
e-mail: alvaro.gonzalezzarza@alum.uca.es

J. Murcia
Instituto Eduardo Torroja, CSIC/IETcc-CSIC, Madrid, Spain
e-mail: murcia@ietcc.csic.es

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_68

In particular, the equilibrium is directly defined by partial differential equations in terms of the shape and the stresses of the membrane, respectively, identified with a negative Gaussian curvature surface and a positive second-order tensor; therefore, we will discuss the following problem: *once the stress tensor is given find the shape of the membrane that balances these same stresses.*

Along with the equilibrium equations, the boundary conditions must be defined; these conditions depend on the used boundary elements. Contrary to the rigid boundaries (see [5]), the shapes of the membrane and its corresponding cable are linked by an unusual relation (see, for instance, [4] and [3]); as a consequence, a not common boundary problem is obtained and a free boundary approach (see [1]) is proposed to solve the general equilibrium equations.

2 Equilibrium Equations of a Membrane

Let us identify the membrane with a surface S with a negative Gaussian curvature; S is the graph of a regular function $z = z(x, y)$, defined in a domain $D \subset \mathbb{R}^2$. If $\boldsymbol{\sigma} := N_{\alpha\beta} = N_{\alpha\beta}(x, y)$, with $\alpha,\ \beta = x$ or y, represents the projected stress tensor of the membrane, the membrane equilibrium equations in terms of $N_{\alpha\beta}$, neglecting its weight and considering no external load, are expressed by (see Fig. 1b and [4] for the details)[1]:

$$\begin{cases} N_{xx,x} + N_{xy,y} = 0 \quad \text{in } D, & \text{(1a)} \\ N_{xy,x} + N_{yy,y} = 0 \quad \text{in } D, & \text{(1b)} \\ N_{xx} z_{,xx} + 2N_{xy} z_{,xy} + N_{yy} z_{,yy} = 0 \quad \text{in } D. & \text{(1c)} \end{cases}$$

In this way, once in system (1) a positive tensor $\boldsymbol{\sigma}$ is fixed the unknown function z has to solve an elliptic equation; this is the problem we want to analyze.

3 Boundary Equilibrium Equations

Once the membrane equilibrium equations are given (system (1)), the problem has to be completed by defining the corresponding boundary conditions on $\Gamma = \partial D$, and in particular on Γ^{r} and Γ^{c} (see Fig. 1c and d).

[1] If $f(x, y)$ is a function, we will write $f_{,x} = \dfrac{\partial f}{\partial x}$, $f_{,y} = \dfrac{\partial f}{\partial y}$, $f_{,xx} = \dfrac{\partial^2 f}{\partial x^2}$ and so on.

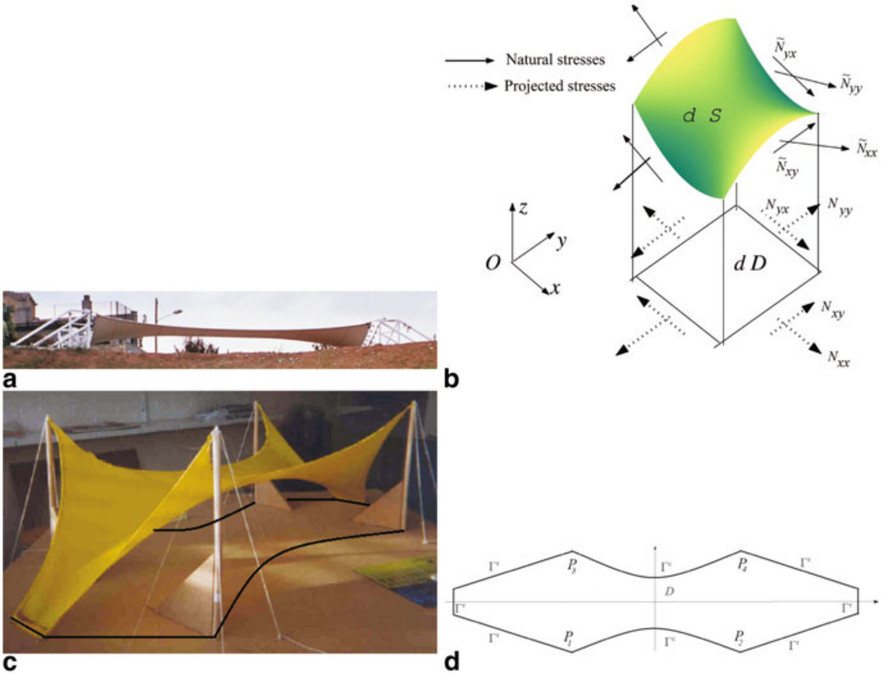


Fig. 1 Main characterizations of a membrane footbridge. We remark that even if the boundary on the model of **c** is totally composed by cable elements (obtained by using guitar strings), in **d** we suppose, without loss of generality, that only the portion corresponding to the *curved black lines* of the same **c** are associated to a cable boundary; it simplifies the formulation of the problem and makes easier the computation of the example in Sect. 5. **a** Picture of a real membrane footbridge built in Spain; see [2] for some interesting technological aspects. **b** Equilibrium equations of a membrane: differential element of a membrane. $\widetilde{N}_{\alpha\beta}$ and $N_{\alpha\beta}$ represent the natural and the projected stresses of the membrane, respectively. **c** A model of a membrane footbridge. The *black line* represents the projection of the boundary of the membrane, and define the domain of the general equilibrium problem (see **d**). **d** Typical projected domain for a footbridge. Γ^{r} is the (projected) rigid boundary and Γ^{c} the (projected) cable boundary; we suppose that Γ^{c} is parameterized by the functions $y = -y(x)$ and $y = y(x)$. Of course $\partial D = \Gamma = \Gamma^{\mathrm{r}} \cup \Gamma^{\mathrm{c}}$

3.1 The Rigid Boundary: Equilibrium Equations

Let us consider the equilibrium on Γ^{r}; as Γ^{r} has bending stiffness, it can assume any shape and moreover, its shape depends neither on the membrane's stresses $\boldsymbol{\sigma}$ nor on its shape z. In this case, the corresponding boundary equilibrium returns the usual Dirichlet condition $z = g$ on Γ^{r}, g being the value of z on the same Γ^{r}, that represents the vertical elevation of the rigid boundary of the membrane.

3.2 The Cable Boundary: Equilibrium Equations

Let us consider the equilibrium on $\varGamma^{\mathrm{c}}$; contrary to the rigid boundary, a cable has no bending stiffness because it works only by means of tensions. Therefore, its shape (necessarily convex) depends on both the membrane's stresses $\boldsymbol{\sigma}$ and its shape z; more exactly, the boundary equilibrium leads to the following *cable-membrane compatibility equation*: $z_{,xx} + 2z_{,xy}y' + z_{,yy}y'^2 = 0 \quad \text{on } \varGamma^{\mathrm{c}}$, $y(x)$ being as in Fig. 1d.

4 Definition and Properties of the Mathematical Problem

4.1 Mathematical Problem: the Equilibrium Equations

Let $N_{\alpha\beta}$ be a positive and symmetric second-order tensor such that both (1a) *and* (1b) *are verified. Find the surface z, such that*

$$\begin{cases} z_{,xx}N_{xx} + 2z_{,xy}N_{xy} + z_{,yy}N_{yy} = \operatorname{div}(\boldsymbol{\sigma} \cdot \nabla z) = 0 \text{ in } D, \\ z = g \text{ on } \varGamma^{\mathrm{r}}, \\ z_{,xx} + 2z_{,xy}y' + z_{,yy}y'^2 = 0 \text{ on } \varGamma^{\mathrm{c}}, \end{cases} \tag{2}$$

where g is a given function on $\varGamma^{\mathrm{r}}$ and $y(x)$, depending on $\boldsymbol{\sigma}$, is the function representing $\varGamma^{\mathrm{c}}$; moreover, $\varGamma^{\mathrm{r}} \cup \varGamma^{\mathrm{c}}$ is the boundary of D (see Fig. 1d).

As shown in [4], one can prove the uniqueness of the solution of system (2); on the other hand, the corresponding existence problem is still an open question.

4.2 Mathematical Problem: A Free Boundary Approach

It can be checked in [3] that problem (2) is equivalent to find z and h, such that

$$\begin{cases} \operatorname{div}(\boldsymbol{\sigma} \cdot \nabla z) = 0 \quad \text{in } D, & \text{(3a)} \\ z = g \text{ on } \varGamma^{\mathrm{r}}, & \text{(3b)} \\ z = h \quad \text{on} \quad \varGamma^{\mathrm{c}}, & \text{(3c)} \\ z_{,y}y'' = h'' \quad \text{on} \quad \varGamma^{\mathrm{c}}. & \text{(3d)} \end{cases}$$

With the aim of solving (3), let us split it into the following problems:

$$A: \begin{cases} \operatorname{div}(\boldsymbol{\sigma} \cdot \nabla z) = 0 \quad \text{in } D, \\ z = g \quad \text{on} \quad \varGamma^{\mathrm{r}}, \\ z = h \quad \text{on} \quad \varGamma^{\mathrm{c}}, \end{cases} \tag{4}$$

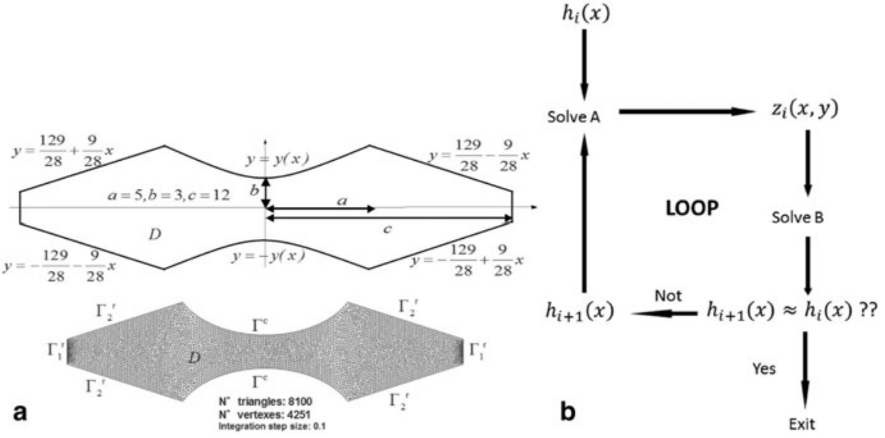


Fig. 2 Some data used in the numerical example and the free boundary scheme. **a** Domain chosen for the analysis of the numerical example (*top*). At the *bottom*, the mesh characterization of D, used to solve problem A); moreover, the step size of problem B is 0.1 (corresponding to 100 nodes). **b** Scheme of the proposed free boundary algorithm; h_0 is the only initial input we need to generate the sequences h_i and z_i

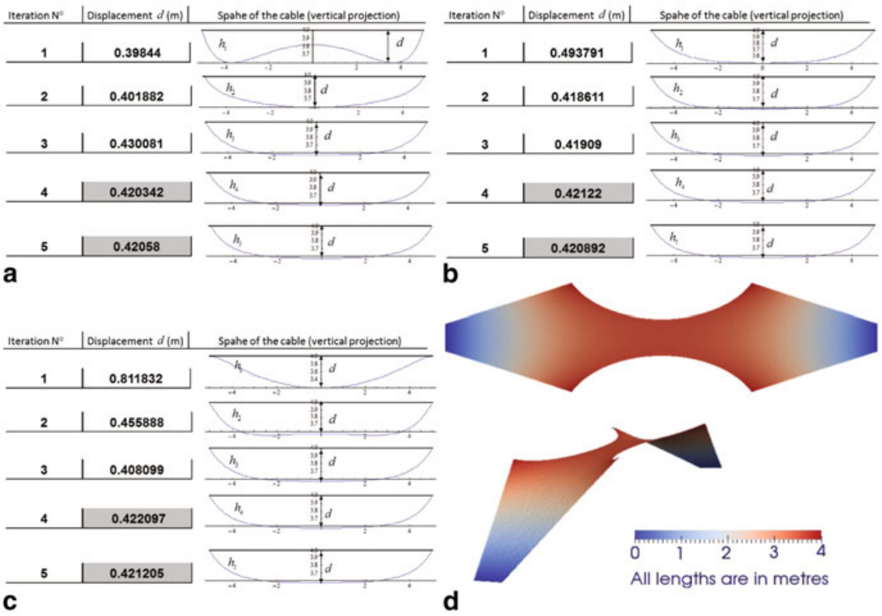
$$B : \begin{cases} h'' = z_{,y} y'' \quad \text{on} \quad \Gamma^{\mathrm{c}}, \\ + \text{ b. c.} \end{cases} \tag{5}$$

We interpret the function $h(x) = z(x, y(x))$ as the free boundary of system (3). In this way, we can consider the iterative method consisting of fixing a function $h_0(x)$ and finding $z_0(x, y)$ solving (4); successively, $z_0(x, y)$ is used as a datum of (5), whose output is another function $h_1(x)$, generally different to the previous one, that represents the new input of (4). This recursive process, schematized in Fig. 2b, generates two sequences $h_i(x)$ and $z_i(x, y)$ whose limits $h(x)$ and $z(x, y)$ make system (3) compatible and determinate.

Remark 1 Of course, it is not possible to fix also the function $h(x)$ in system (3); in fact, in this case (3a–c) would represent a classical Dirichlet problem whose unique solution z would not generally solve also (3d). Consequently the function $h(x)$ has to be an unknown too.

5 A Numerical Example

Let $N_{xx} = 10$ *kN/m*, $N_{xy} = 0$, $N_{yy} = 4$ *kN/m*, and $g = 6 - \frac{26}{43}|x| + \frac{72}{43}|y|$ be the fixed stress tensor and the vertical elevation of the rigid boundary, respectively. As we justified previously, the domain D is represented at the top of Fig. 2a; Γ^{c} is composed by both the graphs of $-y(x)$ and $+y(x)$, with $y(x) = \frac{9}{2} - \sqrt{\frac{49}{4} - \frac{2}{5}x^2}$,



Iteration N°	Displacement d (m)	Spahe of the cable (vertical projection)
1	0.39844	
2	0.401882	
3	0.430081	
4	0.420342	
5	0.42058	

Iteration N°	Displacement d (m)	Spahe of the cable (vertical projection)
1	0.493791	
2	0.418611	
3	0.41909	
4	0.42122	
5	0.420892	

Iteration N°	Displacement d (m)	Spahe of the cable (vertical projection)
1	0.811832	
2	0.455888	
3	0.408099	
4	0.422097	
5	0.421205	

Fig. 3 The free boundary approach: numerical results (all the *lengths* are expressed in *meters*). **a** Case h_0 convex; $h_0 = 1 + \frac{3}{34}(x^2 + y^2)$ on Γ^{c}. Evolution of the vertical projection of h_i and its displacement d. **b** Case h_0 straight; $h_0 = 4$ on Γ^{c}. Evolution of the vertical projection of h_i and its displacement d. **c** Case h_0 concave; $h_0 = 7 - \frac{3}{34}(x^2 + y^2)$ on Γ^{c}. Evolution of the vertical projection of h_i and its displacement d. **d** Numerical result of the final shape of the membrane

and Γ^{r} is defined in the same figure. Therefore, we have to solve these two coupled problems:

$$A : \begin{cases} 10z_{,xx} + 4z_{,yy} = 0 \quad \text{in } D, \\ z = 0 \quad \text{on } \Gamma_1^{\mathrm{r}}, \\ z = g \quad \text{on } \Gamma_2^{\mathrm{r}}, \\ z = h \quad \text{on } \Gamma^{\mathrm{c}}, \end{cases} \qquad B : \begin{cases} h'' = z_{,y} y'', \\ h(\mp a, \mp b) = g(\mp a, \mp b). \end{cases}$$

More exactly, we will apply three times the scheme of Fig. 2b to solve both A and B, analyzing the corresponding solution in terms of the choice of the initial data h_0; moreover, we will use a finite element method to obtain $z_i(x, y)$ (problem A) and a finite difference method to compute $h_i(x, y)$ (problem B). The results summarized in Fig. 3(a–c), show that the procedure is fast and unconditioned; for instance, independently by the initial datum $h_0(x)$, the displacement d of h_i, that model the different shapes of the vertical projections of the cable, approaches to a fixed value when i increases, and in particular at the forth step the final displacement is 0.42 m. On the other hand, in Fig. 3d, we can appreciate the numerical solution of the shape of the surface z representing the membrane.

References

1. Crank, J.: Free and Moving Boundary Problems, Oxford University Press, New York (1984)
2. Murcia, J.: Structural membrane technology for footbridges. Inf. Constr. **59**(507), 21–31 (2007)
3. Viglialoro, G., Murcia, J.: Equilibrium problems in membrane structures with rigid and cable boundaries. Inf. Constr. **63,**(524), 49–57 (2011)
4. Viglialoro, G., Murcia, J.: A singular elliptic problem related to the membrane equilibrium equations. Int. J. Comput. Math. **90**(10), 2185–2196 (2013)
5. Viglialoro, G., Murcia, J., Martínez, F.: Equilibrium problems in membrane structures with rigid boundaries. Inf. Constr. **61**(516), 57–66 (2009)

Approximations to Intractable Spatial Econometric Models and Their Solutions Through Global Optimization

Renata Wachowiak-Smolíková, Mark P. Wachowiak and Jonathan Zimmerling

Abstract Parameter estimation (inverse) problems are ubiquitous in many fields, including spatial econometrics. Global optimization can provide good parameter estimates for many such problems for which traditional, analytic estimation methods fail, or that are otherwise intractable. Stochastic global methods inspired by natural processes have recently gained popularity for difficult optimization problems characterized by imprecise measurements or local optima. In this chapter, one such approach, particle swarm optimization (PSO), is used to estimate parameters of the time series cross-sectional spatiotemporal autoregressive model, a particulary difficult and computationally intensive problem arising in spatial econometrics. Preliminary results are promising, and suggest that stochastic global approaches, and global optimization in general, can successfully address some of these intractable problems.

1 Introduction

Spatial econometrics focuses on how spatial relationships factor into economic models. Spatial factors are geographic and include Euclidean distance, cultural, and political relationships. To be useful, spatial models must adequately fit observations. Accurate parameter estimates not only aid in understanding the effects of spatial relationships but also enhance the model's predictive power. For many models, optimal parameter estimation can be performed by well-known, analytic methods. However, for other very complex models, analytic solutions are not always available, and in

R. Wachowiak-Smolíková (✉) · M. P. Wachowiak · J. Zimmerling
Department of Computer Science and Mathematics, Nipissing University,
North Bay, ON P1B 8L7, Canada
e-mail: renatas@nipissingu.ca

M. P. Wachowiak
e-mail: markw@nipissingu.ca

J. Zimmerling
Stroma Service Consulting Ltd., North Bay, ON, Canada

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_69

some cases no known exact estimator exists. Despite this, there is a continued shift toward increasing complexity in spatial econometric models [6].

In the absence of analytic solutions or tractable approximations, optimization may provide acceptable parameter estimates. Global optimization methods have achieved success in determining optimal (or at least very good) solutions to previously intractable problems. For problems characterized by noisy observations, multiple local optima, high dimensionality or time-complexity, or for those whose derivatives are unavailable or estimates inaccurate, nature-based stochastic techniques, such as genetic algorithms and evolutionary algorithms have been gaining popularity. In this chapter, another such method, particle swarm optimization (PSO), was adapted to estimate parameters of a difficult spatial econometrics model. PSO is relatively straightforward, with few parameters to tune [7]. In contrast to competitive evolutionary methods, the paradigm for PSO is cooperation (e.g., birds flocking while searching for food). PSO has been applied to many parameter estimation problems, including those in econometrics [10]. Furthermore, PSO is inherently parallel, which can increase efficiency for time-intensive problems [9]. Specifically, PSO is applied to the time series cross-sectional spatiotemporal autoregressive model for which no known analytic estimator exists [4]. PSO parameter estimation for this model was previously investigated [11]. In the current chapter, the robustness of PSO in optimizing a computationally intensive cost function involving linear algebra operations is demonstrated with a practical example using imprecise, real data.

2 Time Series Cross-sectional Autoregressive Model

Many spatiotemporal autoregressive models exist (see [11] for a brief review) whose parameters can be estimated with ordinary least squares (OLS), tractable maximum likelihood estimators (MLEs), or the generalized method of moments (GMM). This chapter is primarily concerned with the *time series cross-sectional spatiotemporal autoregressive model* [5], expressed as:

$$\mathbf{y}_t = \beta \mathbf{X}_t + \kappa \mathbf{W} \mathbf{y}_t + \phi \mathbf{y}_{t-1} + \epsilon_t \tag{1}$$

As spatial dependence modeling requires a representation of spatial arrangement, relative spatial positions are represented by $\mathbf{W}$, an $N \times N$ matrix (N is the number of units), with $w_{ij} = 1$ when i and j are neighbors, and 0 otherwise. $\mathbf{W}$ is typically row-standardized so that all rows sum to 1. $\mathbf{X}$ is a matrix of observations, ϵ are temporally independent error terms, and β, κ, and ϕ are parameters to be estimated. The κ $\mathbf{W}\mathbf{y}_t$ term indicates that spatial lag is instantaneous, a condition known as *simultaneity bias*, in which a variable x is dependent upon a variable y that is itself dependent upon x. Common estimators such as OLS and GMM do not perform well in models with simultaneity bias, and hence two MLE approximations were proposed: those of Bhargava and Sargan (BS) and Nerlove and Balestra (NB) [5]. Their parameters are very difficult to estimate, as the derivative of the likelihood function is practically unattainable, and therefore numerical iterative methods are required. Very little work has been done to determine suitable methods for maximizing these likelihoods, as the literature is focused on simpler models, and investigations to determine which of these approximations is preferable, and under what conditions, have been scant [11]. Building upon previous work by the authors on simulated data [11], PSO is employed to estimate the parameters of maximum likelihood functions for these approximations for a model involving real data.

Only a brief summary of the BS and NB approximations can be presented here (see [5] for a full development). For a time series, assume N observations for each time period t, $t = 1, ..., T$. Let $\mathbf{X}$ be the $N \times 1$ column vector of observations. Now, let $\mathbf{B} = \mathbf{I}_N - \kappa \mathbf{W}_t$ and $\mathbf{A} = \phi \mathbf{B}^{-1} - \mathbf{I}_N$. Let $\mathbf{V}_b$ be an $N \times N$ matrix, where

$$\begin{aligned}\mathbf{V}_b = \mathbf{I}_N + \mathbf{A}(\mathbf{I}_N - \phi^2(\mathbf{B}'\mathbf{B})^{-1})^{-1}\mathbf{A}' - \mathbf{A}\phi^{m-1}\mathbf{B}^{-(m-1)} \\ \times (\mathbf{I}_N - \phi^2(\mathbf{B}'\ \mathbf{B})^{-1})^{-1}\phi^{m-1}\mathbf{B}'^{-(m-1)}\mathbf{A}' + \phi^{m-1}\mathbf{B}^{-(m-1)}\phi^{m-1}\ \mathbf{B}'^{-(m-1)}\end{aligned} \tag{2}$$

Let $\mathbf{H}_V$ be the $NT \times NT$ matrix defined as:

$$\mathbf{H}_V \equiv \begin{bmatrix} \mathbf{V} & -\mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ -\mathbf{I} & 2\mathbf{I} & -\mathbf{I} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} & 2\mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & 2\mathbf{I} & -\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & -\mathbf{I} & 2\mathbf{I} \end{bmatrix}$$

where $\mathbf{V}$ is an $N \times N$ matrix, $\mathbf{I} \equiv \mathbf{I}_N$, and $\mathbf{0}$ is an $N \times N$ matrix of zeros. The estimates of the residuals are denoted by the column vector $\hat{\epsilon}$ (see [5] for full calculation details). Further, let $m \geq 2$ denote the number of time units the process has undergone before observation began, and define $\mathbf{V}_{BS} \equiv \mathbf{I}_N + \mathbf{V}_b$. Given these defintions, the BS maximum likelihood approximation is:

$$\begin{aligned}\ln \mathcal{L} = -\frac{NT}{2}\ln(2\pi\sigma^2) + T\sum_{i=1}^{N}\ln(1-\kappa\mathbf{w}_i) - 2\ln\big[1 - T + \frac{2T(1-\kappa\mathbf{w}_i)}{1-\kappa\mathbf{w}_i+\phi} \\ \times \big(1 + (\frac{\phi}{1-\kappa\mathbf{w}_i})^{2m-1} + T\phi^2\big)\big] - \frac{1}{2\sigma^2}\Delta\epsilon'\mathbf{H}^{-1}_{V_{NB}}\Delta\epsilon\end{aligned} \tag{3}$$

Now let the covariance matrix of $\Delta\mathbf{X}$ be denoted as $\Sigma_{\Delta\mathbf{X}}$, and define $\mathbf{V}_{NB} \equiv \mathbf{V}_b + \sigma^{-2}Var[\mathbf{X}]$, where $Var[\mathbf{X}]$ is an estimate of the variance of $\mathbf{X}$. The NB maximum likelihood approximation is:

$$
\begin{aligned}
\ln \mathcal{L} = & -\frac{NT}{2}\ln(2\pi\sigma^2) + T\sum_{i=1}^{N}\ln(1-\kappa\mathbf{w}_i) - \frac{1}{2\sigma^2}\Delta\epsilon'\mathbf{H}_{V_{NB}}^{-1}\Delta\epsilon \\
& -\frac{1}{2}\sum_{i=1}^{N}\ln\Bigg[1 - T + \frac{2T(1-\kappa\mathbf{w}_i)}{(1-\kappa\mathbf{w}_i+\phi)}\Big(1+\Big(\frac{\phi}{1-\kappa\mathbf{w}_i}\Big)^{2m-1}\Big) \qquad (4) \\
& + T\frac{\beta\Sigma_{\Delta\mathbf{x}}\beta(1-\kappa\mathbf{w}_i+\phi)^2}{\sigma^2(1-\kappa\mathbf{w}_i)^2}\Big(1-\Big(\frac{\phi}{1-\kappa\mathbf{w}_i}\Big)^m\Big)^2\Bigg]
\end{aligned}
$$

In both approximations, σ can be replaced by its first-order maximizing condition $\hat{\sigma}^2 = (NT)^{-1}\Delta\hat{\epsilon}'\mathbf{H}_{V_b}^{-1}\Delta\hat{\epsilon}$ [5].

3 Methods

3.1 Particle Swarm Optimization

PSO is an iterative, stochastic global optimization method to determine $\mathbf{x}^* = \arg\min f(\mathbf{x})\forall\,\mathbf{x} \in \Omega^D$, where Ω^D is the D-dimensional search space [7]. D-D positions of N independent "particles" search Ω^D for an optimum solution. In each iteration, the positions of the particles are updated from search results. The best position found by each particle $\mathbf{x}_i$ is its "personal best," $\mathbf{x}_i^{best}$. $\mathbf{g}^{best}$ represents the globally best position in the entire population. Both values are updated during the search. For each particle $i = 1, ..., N$, $\mathbf{v}_i$ at time $t+1$ is updated as:

$$\mathbf{v}_i(t+1) = \chi[\mathbf{v}_i(t) + C_1 r_1(t)(\mathbf{x}_i^{best}(t) - \mathbf{x}_i(t)) + C_2 r_2(t)(\mathbf{g}^{best}(t) - \mathbf{x}_i(t))], \qquad (5)$$

C_1 and C_2 are local and global acceleration constants, respectively, where $C_1 + C_2 \approx 4$, $r_1, r_2 \sim U(0,1)$ are stochastic factors to maintain population diversity, and χ is a constriction factor to decrease the effect of particles' previous velocities, and is generally set to $\chi \approx 0.729$ [2]. The position of each particle $\mathbf{x}_i$ is then updated as $\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t)$. Many application-specific PSO variations have been proposed [3, 8]. Here, PSO is used to maximize the BS and NB MLEs (Eqs. 3, 4).

3.2 Model Based on Real Data

Pooled panel data (observations of multiple phenomena obtained over multiple time periods) were obtained on per capita income and cigarette sales from $N = 46$ states spanning 30 years ($T = 29$) (http://www.wiley.co.uk/baltagi) [1]. A three-parameter model was constructed for cigarette sales in those states [1]: $y_{i,t} = x_{i,t}\beta + \phi y_{i,t-1} + \kappa\mathbf{W}y_{i,t} + \epsilon_{i,t}$, where $x_{i,t}$ is the per capita disposable income. $\mathbf{W}$ is an $N \times N$ binary contiguity matrix with columns $\mathbf{w}_i$, $i = 1, ..., N$. $w_{i,j} = 1$ if states i and j share a border, or 0 otherwise, yielding a symmetric matrix that is invariant over the time. Each row of $\mathbf{W}$ was normalized unit sum, allowing the search to be confined to $|\hat{\kappa}| < 1$. The $\mathbf{y}_t$ are dependent upon the disposable income of the residents of this state, the previous year's value of $\mathbf{y}$ in that state, and the current value of $\mathbf{y}$ in all other states with which it shares a border. For the experiment, $y_{i,0} \sim \mathbf{U}(0, 4)$. Additionally, "ground-truth" values were empirically calculated based on extensive experimentation to yield the best fit of the data to the model: $\beta = -1.8$, $\phi = 0.8$, and $\kappa = 0.5$. It was also assumed that the spatial component occurs away from the per capita income variable. The income of workers in one state will be linked to the income of workers in others, but as these are exogenous variables, spatial effects need not be considered. No data were recorded for Alaska, Hawaii, Colorado, or North Carolina. Parameters are therefore estimated for a "real/simulated" hybrid, as real-world economic panel data for Eq. 1 is challenging to produce; although, the model is very useful [5], real data sets are not readily available.

PSO was run with 500 particles initially distributed randomly throughout the 3-D search space. For each approximation, PSO ran for a 1000 iterations, or until the average distance between the particles and the centroid was less than 0.001. The programs for the experiments were written in MATLAB (The Mathworks, Natick, MA). Coefficients of determination ($R^2 \in [0, 1]$) were calculated by comparing the ground truth data with those predicted by the model with the estimated parameters. A trial was considered a "success" if R^2 is close to 1, where $R^2 = 1$ indicates that the parameterized model perfectly fits the observed data.

4 Results

The success of the optimization is principally reflected in the R^2 values, as the parameters are applied to "real" disposable incomes to predict "real" cigarette sales. Although, the "ground-truth" parameter values are themselves empirical estimates obtained to best fit the selected model, comparison of these values with those obtained from the optimization also provide some indication of PSO performance. The estimates $\hat{\beta}$, $\hat{\kappa}$, and $\hat{\phi}$ from the PSO experiments on the real/simulated cigarette data model are shown in Table 1. Although promising R^2 values are obtained, there is a significant bias in the estimation of β, as shown in Table 1. The BS approximation generally outperformed the NB approximation, as was observed with previous experiments on simulated data [11]. The better R^2 for BS is likely due to NB relying on the sample covariance matrix of $\mathbf{X}$, which is prone to computation inaccuracies. It is also assumed that $\mathbf{X}$ is a stationary process for each state. In reality, per capita income will likely change over time.

The relatively high bias values, especially for $\hat{\beta}$, can be attributed to the global PSO algorithm becoming entrapped in a local minimum. However, as evidenced by $R^2 > 0.7$ for both approximations, the parameter estimates returned by PSO were good, although (from the observed biases) not optimal. The complexity of the cost function (Eqs. 3 and 4), the difficult landscape of the cost function, error

Table 1 Results for the Nerlove–Balestra (NB) and Bhargara–Sargan (BS) approximations

	$\hat{\beta}$	Bias $\hat{\beta}$	$\hat{\kappa}$	Bias $\hat{\kappa}$	$\hat{\phi}$	Bias $\hat{\phi}$	R^2
NB	– 3.7031	– 1.9031	0.3002	– 0.1998	1.1196	0.3196	0.7205
BS	– 2.4560	– 0.6516	0.1727	– 0.3273	1.3236	0.5262	0.9597

propagation through the lengthy $f(\mathbf{x})$ computation, and the imperfect representation of the demand data by the cross-sectional model all contributed to this nonoptimality [11].

5 Conclusion

Experimental results confirm that PSO can be used to accurately estimate parameters of complex spatial econometrics models with real data. Each experiment required about 7 h on a 2.8 GHz Intel®Xeon®CPU. Efficiency can be greatly improved by parallelization on multicore CPUs or graphics processors, as PSO is inherently parallel (each particle i evaluates $f(\mathbf{x}_i)$ simultaneously in each iteration) [9]. In future work, in addition to parallelization, enhancements to PSO and combination with deterministic local methods will reduce entrapment in local minima and will better refine final estimates.

Global optimization approaches for parameter estimation of large, complex models, including the cross-sectional autoregressive model presented here, appear to be very good options. For even simple spatial econometrics models, global optimization should at least be considered, as local approaches fail, and because the good initial guesses required for MLE and GMM application are often unavailable.

Acknowledgments The authors thank Mr. Devin Rotondo for his assistance. M.P.W. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC 386586-2011).

References

1. Baltagi, B.H., Griffin, J.M., Xiong, W.: To pool or not to pool: homogeneous versus heterogenous estimators applied to cigarette demand. Rev. Econ. Stat. **82,** 117–126 (2000)
2. Banks, A., Vincent, J., Anyakoha, C.: A review of particle swarm optimization. Part i: background and development. Nat. Comput. **6**(4), 467–484 (2007)
3. Banks, A., Vincent, J., Anyakoha, C.: A review of particle swarm optimization. Part ii: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications. Nat. Comput. **7**(1), 109–124 (2008)
4. Beck, N., Gleditsch, K.S., Beardsley, K.: Space is more than geography: using spatial econometrics in the study of political economy. Int. Stud. Q. **50**(1), 27–44 (2006)
5. Elhorst, J.P.: Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. Geogr. Anal. **37**(1), 85–106 (2005)
6. Elhorst, J.P.: Applied spatial econometrics: raising the bar. Spat. Econ. Anal. **5**(1), 9–28 (2010)

7. Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann, San Francisco (2001)
8. Schutte, J.F., Reinbolt, J.A., Fregly, B.J., Haftka, R.T., George, A.D.: Parallel global optimization with the particle swarm algorithm. Int. J. Numer. Methods Eng. **61**(13), 2296–2315 (2004)
9. Wachowiak, M.P., Lambe Foster, A.E.: Gpu-based asynchronous global optimization with particle swarm. J. Phys. Conf. Ser. **385**(1) (2012)
10. Wachowiak, M., Wachowiak-Smolíková, R., Smolik, D.: Parameter estimation of nonlinear econometric models using particle swarm optimization. Cent. Eur. Rev. Econ. Issues **13,** 193–199 (2010)
11. Wachowiak, M., Wachowiak-Smolíková, R., Zimmerling, J.: The viability of global optimization for parameter estimation in spatial econometrics models. In: European Regional Science Association: Proceedings of the 53rd ERSA Congress, vol. 12 (2012)

Application of Advanced Diagonalization Methods to Quantum Spin Systems

Jie Yu Wang and Ralf Meyer

Abstract In this work, the Block Davidson and the residual minimization-direct inversion in the iterative subspace (RMM-DIIS) algorithms are used to diagonalize the Hamilton matrices arising from antiferromagnetic spin-$\frac{1}{2}$ Heisenberg models. The results show that both algorithms find reliably the lowest eigenvalues but the computational costs are smaller for the RMM-DIIS method. In addition to this, the authors show that the new Intel Xeon Phi coprocessor can be used efficiently for this type of problems.

1 Introduction

Quantum spin models play an important role in theoretical condensed matter physics and quantum information theory. A numerical technique that is frequently used to study quantum spin systems is exact diagonalization. In this approach, numerical methods are used to find the lowest eigenvalues and corresponding eigenvectors of the Hamilton matrix that describes the quantum system. The computational problem is thus to determine the lowest eigenpairs of an extremely large, sparse matrix. An overview of the exact diagonalization technique is given in [12].

Although many sophisticated iterative techniques for the determination of a small number of lowest eigenpairs can be found in the literature, most exact diagonalization studies of quantum spin systems have employed the Lanczos algorithm. In contrast to this, other methods have been applied very successfully to the similar quantum mechanics problem of electronic structure calculations. The well-known VASP code [4], for example, provides an implementation of the Block Davidson method [3, 5, 7] as well as the residual minimization-direct inversion in the iterative subspace algorithm (RMM-DIIS) [9, 10, 13].

R. Meyer (✉) · J. Y. Wang
Department of Mathematics and Computer Science, Laurentian University,
935 Ramsey Lake Road, Sudbury, ON P3E 2C6, Canada
e-mail: rmeyer@cs.laurentian.ca

J. Y. Wang
e-mail: jy2_wang@laurentian.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_70

In this work, we study the efficiency of the Block Davidson and RMM-DIIS method when applied to quantum spin models like the spin-$\frac{1}{2}$ Heisenberg chain, ladder, and dimerized ladder [1, 6, 8, 11]. Results are presented that allow a comparison of the algorithms based on the number of iterations to achieve convergence and the required computational time. An important aspect of state-of-the-art scientific calculations is the parallel performance on current high-performance computing equipment. To this end we have tested our RMM-DIIS implementation on an Intel Xeon Phi coprocessor and show the resulting parallel speedups.

2 Computational Methods

The Davidson method was introduced by E. R. Davidson as a method for the solution of large eigenvalue problems in quantum chemistry [3]. Subsequently, it was improved by Liu [5] and Murray et al. [7]. In this work we use the generalized (block) formulation by Crouzeix et al. [2]. Each iteration of the method consists of a Rayleigh–Ritz step followed by the application of a preconditioner and a modified Gram–Schmidt orthogonalization procedure. We follow the original Davidson method and use a simple diagonal preconditioner of the form $C_{k,i} = (\lambda_{k,i} I - D)^{-1}$ where $\lambda_{k,i}$ is the approximation of the ith eigenvalue at the kth step and D is the diagonal part of the matrix A whose lowest eigenvalues are sought.

The RMM-DIIS method was originally proposed by Pulay [9]. Wood and Zunger were the first to apply the method to the eigenvalue problems in electronic structure calculations [13]. An advantage of the RMM-DIIS method is that it does not require a computationally expensive explicit orthogonalization step at each iteration. Since the method is based on the minimization of the residual vector, the trial vectors converge to the eigenvectors with eigenvalues closest to the current trial eigenvalues. In our implementation we periodically perform a Ritz projection of the lowest eigenvectors in order to separate the eigenvectors of degenerate eigenvalues.

Since the RMM-DIIS method converges toward eigenpairs close to the current trial eigenvalues, some care has to be taken at the beginning of the computations when there are no good trial eigenvalues. We tested two approaches: In the first method we started the calculation with the Block Davidson method and switched to the RMM-DIIS method after 60 iterations or when the norms of all residual vectors were below 5×10^{-2}. In the second approach we immediately started with the RMM-DIIS method. However, in order to avoid convergence toward higher eigenvalues, we included all intermediate steps in the periodic Ritz projection (we call this variant of the algorithm maximum-V). This enlarges the search space of the Ritz projection and leads, in all of our calculations, to a convergence toward the lowest eigenvalues. As in the first approach, we switched to the normal variant of the RMM-DIIS algorithm (which we call minimum-V) after 60 iterations or when the norms of all residual vector were below 5×10^{-2}.

Our implementations of the Block Davidson and RMM-DIIS algorithms can be run in parallel on shared-memory computers. To this end, the program makes heavy use of Intel's Math Kernel Library (MKL). In particular, multiplications of the sparse

Hamilton matrix with a vector are performed with the help of the MKL. Additionally, we used OpenMP in order to parallelize some vector operations not performed with the MKL. All calculations described in this work were performed on an Intel Xeon Phi coprocessor 5110P. This coprocessor integrates 60 core with 4 hardware threads per core.

3 Results

To compare the efficiency of the Block Davidson and RMM-DIIS method for the exact diagonalization of quantum spin models, we have used different types of antiferromagnetic Heisenberg models with spin $S = \frac{1}{2}$ [8]. Each of these models represents a system of N interacting spins that can only take the values $\pm\frac{1}{2}$. The dimension of the state space of these models is 2^N and they are characterized by a Hamilton matrix of dimension $2^N \times 2^N$. The symmetries of the Heisenberg models allow, however, a block-diagonalization of the matrix. In this work, we exploit the conservation of the z-component of the spin to reduce the dimension of the problem to $\left(\frac{N}{(N/2)(N-N/2)}\right)$. The dimension of the eigenvalue problem is thus 12,870, 48,620, 184,756, 705,432, 2,704,155, and 10,400,600 for $N = 16$, 18, 20, 22, 24, and 26 spins, respectively.

The first models that we used for our tests are a simple Heisenberg chain [8] and a two-leg ladder [1] with the ratio between the coupling along the rungs and the coupling along the legs $J_\perp/J_\| = 1$. In addition to this, we employed dimerized two-leg ladder models [6, 11] with a dimerization parameter $\gamma = 0.5$ (cf. [6]) and $J_\perp/J_\| = 1$ and 0.5 (Ladder 215 and 255). Finally, we applied the algorithms to dimerized three-leg ladder systems with $\gamma = 0.5$ and $J_\perp/J_\| = 1$ (Ladder 315). For the dimerized ladder systems, we considered both the columnar and the staggered dimerization patterns.

Tables 1, 2, and 3 show the number of iterations and the execution time of the Block Davidson method as well as both variants of the RMM-DIIS methods for three of our benchmark models with $N = 24$. Iterations were stopped when the norms of all residual vectors were below 5×10^{-4}. The results show that except for the chain model, the number of iterations to reach convergence from random start vectors is not very different between the three methods. However, the execution time clearly favors the RMM-DIIS method due to the lower number of expensive orthogonalization steps.

In Fig. 1, we show the parallel speedup obtained with the RMM-DIIS method on the Intel Xeon Phi coprocessor for the dimerized two-leg Heisenberg ladder 215 model for various system sizes N. We tested two settings of the environment variable KMP_THREAD_AFFINITY. This variable affects the binding of the threads to processor cores for OpenMP and the MKL on the coprocessor. If set to "scatter" (left panel) the runtime system distributes the threads over as many cores as possible whereas "compact" (right panel) uses as few cores as possible. The figure shows that

Table 1 Heisenberg chain, $N = 24$, lowest six eigenvalues

Method	Block size	Iterations	Convergent	Time (s)	Memory (MB)
Block Davidson	4	125	Yes	378.5	17878
Block Davidson-	4	39	Yes	118.7	17878
RMM-DIIS minimum V	4	69(23)	Yes	131.8	17733
RMM-DIIS maximum V	4	51(17)	Yes	14.3	18208
RMM-DIIS minimum V	4	30(10)	Yes	20.2	17733

Table 2 Heisenberg ladder, $N = 24$, lowest eight eigenvalues

Method	Block size	Iteration	Convergent	Time (s)	Memory (MB)
Block Davidson	4	169	Yes	743.2	18376
Block Davidson-	4	60	No	263.2	18376
RMM-DIIS minimum V	4	72(24)	Yes	300.1	18149
RMM-DIIS maximum V	4	60(20)	No	28.6	18810
RMM-DIIS minimum V	4	120(40)	Yes	38.8	18149

Table 3 Dimerized two-leg Heisenberg ladder (Ladder 215), $N = 24$, lowest eight eigenvalues

Method	Block size	Iteration	Convergent	Time (s)	Memory MB
Block Davidson	4	125	Yes	504.1	18392
Block Davidson-	4	47	Yes	194.1	18392
RMM-DIIS minimum V	4	72(24)	Yes	214.2	18133
RMM-DIIS maximum V	4	60(20)	No	25.7	18814
RMM-DIIS minimum V	4	87(29)	Yes	49.9	18133

for the small systems with $N = 16$, 18, and 20, the coprocessor is inefficient with maximum speedups far below what would be expected from a 60 core device. For $N = 22$, a maximum speedup of 30 is reached at 120 threads whereas for $N = 24$ and 26, the speedup rises continuously until $p = 240$ threads. The maximum speedups for $N = 24$ and 26 range from 50 to 60. It can also be seen that for the larger systems compact core binding gives better speedups than scattered.

In Table 4, we compare the parallel speedups obtained for all our benchmark models with $N = 24$ on the Xeon Phi coprocessor for different numbers of threads. As in Fig. 1 we have considered both types of thread binding. For thread numbers p up to 60, the speedups are similar for all models and correspond to a parallel efficiency of about 33 %. As the number of threads p goes beyond 60 (i.e., more than one thread per core), the speedup continues to increase but differences between the models appear. This is best seen at $p = 240$. The highest speedup is obtained by the three-leg dimerized ladder systems. All two-leg ladder models have slightly lower speedups followed by the chain system. The data show that the speedup is determined by the number of legs of the model. All two-leg ladders have (nearly)

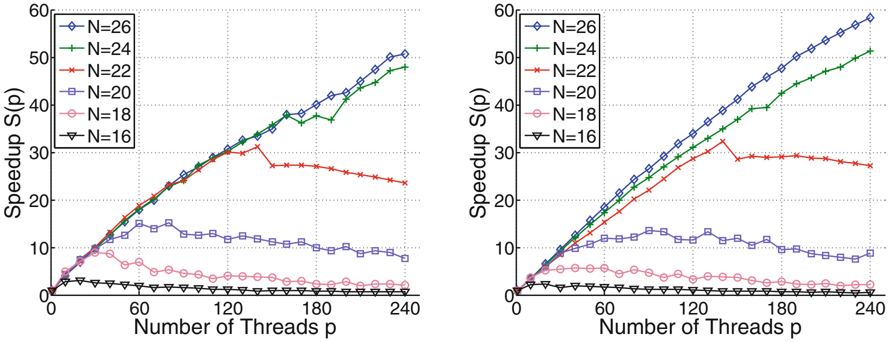


Fig. 1 Parallel speedup of the dimerized Heisenberg ladder model (Ladder 215) with $N = 16$, 18, 20, 22, 24, and 26 on an Intel Xeon Phi coprocessor with KMP_THREAD_AFFINITY set to scatter (*left*) and compact (*right*)

Table 4 Parallel speedup of the RMM-DIIS algorithm on an Intel Xeon Phi coprocessor for different models with $N = 24$

Type	Thread	1	30	60	90	120	150	180	210	240
Chain	Scatter	1.0	10.3	18.1	23.1	29.7	33.6	36.2	40.4	43.9
	Compact	1.0	8.7	16.2	23.0	28.8	33.8	38.8	43.0	46.3
Ladder	Scatter	1.0	10.3	18.6	24.6	31.3	36.5	38.7	42.8	47.8
	Compact	1.0	9.2	17.2	24.7	31.0	36.9	42.3	47.1	51.3
Ladder 215	Scatter	1.0	10.1	18.5	25.9	31.1	35.1	38.1	43.9	48.0
Columnar	Compact	1.0	9.3	17.2	24.7	31.0	36.8	42.4	47.2	51.4
Ladder 215	Scatter	1.0	10.0	18.1	24.0	30.2	35.8	37.7	43.6	47.9
Staggered	Compact	1.0	9.3	17.3	24.7	31.0	37.0	42.4	47.1	51.3
Ladder 255	Scatter	1.0	9.8	17.6	25.8	30.4	35.1	38.5	43.7	47.8
Columnar	Compact	1.0	9.3	17.3	24.7	31.0	36.8	42.3	47.1	51.3
Ladder 255	Scatter	1.0	9.9	18.5	25.2	30.3	34.9	37.0	43.7	47.5
Staggered	Compact	1.0	9.3	17.3	24.7	31.0	36.9	42.3	47.0	51.3
Ladder 315	Scatter	1.0	10.2	18.1	24.8	31.4	35.2	39.3	43.6	48.5
Columnar	Compact	1.0	9.0	17.6	25.1	31.6	37.4	43.4	48.3	52.6
Ladder 315	Scatter	1.0	9.9	19.4	25.6	31.0	35.8	39.1	43.8	48.6
Staggered	Compact	1.0	9.4	17.5	25.1	31.6	37.4	43.3	48.2	52.7

the same speedups independent of the other parameters. The same is true for the three-leg models. This indicates that the speedup is determined by the sparsity of the matrix. The number of nonzero entries per row is 13, 19, and 21 for the chain, two-leg and three-leg models, respectively.

4 Summary and Conclusions

In this work, we study the efficiency of the Block Davidson method and the RMM-DIIS algorithm for the exact diagonalization of quantum spin models. The results show clearly that the RMM-DIIS algorithm performs better for these systems since it does not require explicit orthogonalization of the trial eigenvectors at each iteration.

On an Intel Xeon Phi coprocessor (60 cores, 4 hardware threads per core) we obtained good parallel speedups of more than 50 for larger systems with at least $N = 24$ spins. The highest speedup was obtained on this machine for the three-leg ladder models. The results suggest the parallel performance of the coprocessor in these calculations is limited by the extreme sparsity of the matrices of our models. We expect therefore even higher speedups for models with a less sparse matrix.

Acknowledgments This work has been financially supported by Laurentian University and the *Natural Science and Engineering Research Council of Canada* (NSERC). We gratefully acknowledge generous allocation of computer time on the facilities of the *Shared Hierarchical Academic Research Network* (SHARCNET) and Compute/Calcul Canada.

References

1. Barnes, T., Dagotto, E., Riera, J., Swanson, E.S.: Excitation spectrum of Heisenberg spin ladders. Phys. Rev. B **47**, 3196–3203 (1993)
2. Crouzeix, M., Phillipe, B., Sadkane, M.: The Davidson method. SIAM J. Sci. Comput. **15**(1), 62–76 (1994)
3. Davidson, E.R.: The iterative calculation of a few lowest eigenvalues and corresponding eigenvectors of large symmetric matrices. Comput. Phys. **17**, 87–94 (1975)
4. Kresse, G., Furthmüller, J.: Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput. Mater. Sci. **6**, 15–50 (1996)
5. Liu, B.: The simultaneous expansion for the solution of several of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. In: Numerical Algorithms in Chemistry: Algebraic Method, vol. LBL-8158, pp. 49–53. Lawrence Berkeley Lab., CA (1978)
6. Martín-Delgado, M.A., Shankar, R., Sierra, G.: Phase transitions in staggered spin ladders. Phys. Rev. Lett. **77**, 3443–3446 (1996)
7. Murray, C.W., Racine, S.C., Davidson, E.R.: Improved algorithms for the lowest few eigenvalues and associated eigenvectors of large matrices. J. Comput. Phys. **103**, 382–389 (1992)
8. Parkinson, J.B., Farnell, D.J.J.: An Introduction to Quantum Spin Systems. Lecture Notes in Physics, vol. 816. Springer-Verlag, Berlin (2010)
9. Pulay, P.: Convergence acceleration of iterative sequences. The case of SCF iteration. Chem. Phys. Lett. **73**, 393–398 (1980)
10. Thijssen, J.: Computational Physics, 2nd edn. Cambridge University Press, Cambrige (2010)
11. Totsuka, K., Suzuki, M.: The spin-1/2 Heisenberg spin ladder with bond alternation. J. Phys. Condens. Matter **7**, 6079–6096 (1995)
12. Weiße, A., Fehske, H.: Exact diagonalization techniques. In: Computational Many-Particle Physics, Lecture Notes in Physics, vol. 739, pp. 529–544. Springer-Verlag, Berlin (2008)
13. Wood, D.M., Zunger, A.: A new method for diagonalising large matrices. J. Phys. A Math. Gen. **18**, 1343–1359 (1985)

The Effects of Body Fluid on Cheyne–Stokes Respiration

Marianne Wilcox and Allan R. Willms

Abstract A compartmental model of the human circulatory system that illustrates Cheyne–Stokes respiration (CSR) is presented. Clinical evidence suggests that patient body position can influence the likelihood of experiencing CSR, and this model is analyzed to see if blood fluid shifts associated with body position could be the means of this influence. It is shown that lying down causes a shift in the location of the Hopf bifurcation curve associated with the onset of CSR, making it more likely.

1 Introduction

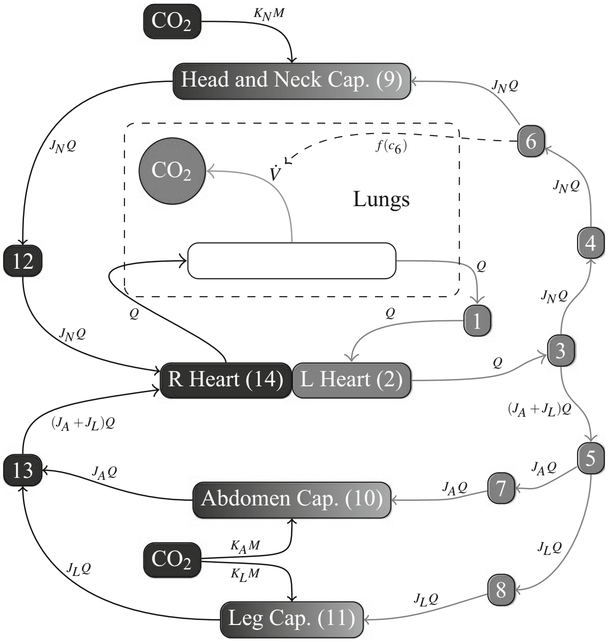
Cheyne–Stokes respiration (CSR) is a form of central sleep apnea (CSA) characterized by an abnormal breathing pattern cycling between apnea (temporary breathing cessation) and hyperpnea (rapid and deep breathing). In contrast to obstructive sleep apneas (OSAs), which are caused by pharynx collapse to varying degrees, CSAs are neurological in origin. In particular, CSR is associated with the brain's monitoring of carbon dioxide (CO_2) levels in the blood. The fundamental features causing the periodic breathing associated with CSR are the sensitivity of the chemoreceptors in the neck to CO_2 levels and the delay in the signal from the lungs to the receptors. CSR is analogous to trying to get warm water from the end of a long hose by adjusting a very responsive hot water faucet. The water is too cold so you open the hot faucet more, but then it is too hot so you close the faucet some, and so on. CSR is common among patients with congestive heart failure and is correlated with increased mortality [7].

Clinical evidence suggests that the likelihood of experiencing CSR is dependent on body position [9], being more likely to occur in patients who are in a supine rather than a sitting or upright position. It is hypothesized that blood volume shifts in response to body position are responsible for this increased likelihood [10].

M. Wilcox (✉) · A. R. Willms
University of Guelph, Guelph, ON, N1G 2W1, Canada
e-mail: wilcoxm@uoguelph.ca

A. R. Willms
e-mail: awillms@uoguelph.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_71



compartments	
1	Pulmonary veins
2	Left heart
3	Aorta
4	Carotid arteries
5	Other main arteries
6	Arterioles neck and head
7	Arterioles abdomen and thorax
8	Arterioles legs
9	Capillaries neck and head
10	Capillaries abdomen and thorax
11	Capillaries legs
12	Jugular vein
13	Systemic veins
14	Right heart and pulmonary arteries
15	Pulmonary capillaries

Fig. 1 Circulatory system compartments and CO$_2$ exchange

This chapter develops and analyzes a compartment model of the human circulatory system to investigate whether blood volume shifts can have the effect of increasing the likelihood of CSR [12]. The model is a generalization of one studied by Langford and others [1, 2], where a number of the compartments have been further divided to allow separation of the legs, abdomen and thorax, and neck and head.

2 Model

The model has 15 compartments with blood flowing between them (see Fig. 1).

The volume of compartment i (in L) will be denoted v_i. Let c_i denote the concentration of CO$_2$ in compartment i, measured in mL of CO$_2$ at standard temperature and pressure, dry (STPD) per L of blood. The overall blood flow rate Q (L/min) is divided into three portions corresponding to the flow to the legs (J_L), the abdomen and thorax (J_A), and the neck and head (J_N), so $J_L + J_A + J_N = 1$. Similarly, the overall metabolic rate M (mL of CO$_2$ at STPD/min) is divided into these three main body areas with the fractions K_L, K_A, and K_N, with $K_L + K_A + K_N = 1$. Metabolism adds CO$_2$ to the blood in the systemic capillaries (compartments 9, 10, and 11) and CO$_2$ is removed from the blood in the pulmonary capillaries (compartment 15). The rate of removal of CO$_2$ is governed by the ventilation rate $\dot{V}$ (L/min), which in turn is controlled by the CO$_2$ sensors located in the neck arterioles (compartment 6). This control function is modeled as the Hill function

$$\dot{V} = f(c_6) = \overline{\dot{V}} \frac{2c_6^n}{a^n + c_6^n}, \tag{1}$$

where $\overline{\dot{V}}$ is the half-maximal ventilation rate ($\approx$ 5 L/min), $n > 0$ is the Hill function power, and a is the CO_2 concentration in the neck arterioles at which the half-maximal ventilation rate occurs.

The governing equations for the model are then

$$\begin{aligned}
v_1 \frac{dc_1}{dt} &= Q(c_{15} - c_1), & v_9 \frac{dc_9}{dt} &= J_N Q(c_6 - c_9) + K_N M,\\
v_2 \frac{dc_2}{dt} &= Q(c_1 - c_2), & v_{10} \frac{dc_{10}}{dt} &= J_A Q(c_7 - c_{10}) + K_A M,\\
v_3 \frac{dc_3}{dt} &= Q(c_2 - c_3), & v_{11} \frac{dc_{11}}{dt} &= L_L Q(c_7 - c_{11}) + K_L M,\\
v_4 \frac{dc_4}{dt} &= J_N Q(c_3 - c_4), & v_{12} \frac{dc_{12}}{dt} &= J_N Q(c_9 - c_{12}),\\
v_5 \frac{dc_5}{dt} &= (J_L + J_A) Q(c_3 - c_5), & v_{13} \frac{dc_{13}}{dt} &= Q(J_A c_{10} + J_L c_{11} - (J_L + J_A) c_{13}),\\
v_6 \frac{dc_6}{dt} &= J_N Q(c_4 - c_6), & v_{14} \frac{dc_{14}}{dt} &= Q(J_N c_{12} + (J_L + J_A) c_{13} - c_{14}),\\
v_7 \frac{dc_7}{dt} &= J_A Q(c_5 - c_7), & v_{15} \frac{dc_{15}}{dt} &= Q(c_{14} - c_{15}) - D \frac{\dot{V}}{V_G} (p_A - p_I),\\
v_8 \frac{dc_8}{dt} &= J_L Q(c_5 - c_8).
\end{aligned}$$

In the above, V_G is the alveolar gas volume, p_A is the alveolar partial pressure of CO_2, p_I is the partial pressure of CO_2 of inspired air, and D is a dissociation conversion factor for CO_2 given by

$$D = \frac{V_G T_S}{P_S T_B},$$

where T_S and T_B are the standard and body temperatures, and P_S is standard pressure. An empirical relationship,

$$p_A = \frac{c_{15}}{8} - 20, \tag{2}$$

relates the alveolar partial pressure of CO_2 with the concentration of CO_2 in the pulmonary capillaries [2].

Base-line values for the volumes of the compartments and the metabolic and blood flow proportions were taken or extrapolated from various sources [3–6, 11]. The total volume of blood was normalized to 5 L. These values and the values of other parameters in the model are given in Table 1.

Table 1 Model parameter values

Symbol	Name	Value	Units
v_1	Volume of pulmonary veins	0.25	L
v_2	Volume of left heart	0.2	L
v_3	Volume of aorta	0.1	L
v_4	Volume of carotid arteries	0.025	L
v_5	Volume of other main arteries	0.375	L
v_6	Volume of arterioles neck and head	0.01	L
v_7	Volume of arterioles abdomen and thorax	0.0275	L
v_8	Volume of arterioles legs	0.0125	L
v_9	Volume of capillaries neck and head	0.025	L
v_{10}	Volume of capillaries abdomen and thorax	0.175	L
v_{11}	Volume of capillaries legs	0.05	L
v_{12}	Volume of jugular vein	0.2	L
v_{13}	Volume of systemic veins	3.05	L
v_{14}	Volume of right heart and pulmonary arteries	0.4	L
v_{15}	Volume of pulmonary capillaries	0.1	L
Q	Blood flow rate	5	L/min
J_L	Blood flow fraction legs	0.15	
J_A	Blood flow fraction abdomen and thorax	0.69	
J_N	Blood flow fraction neck and head	0.16	
M	Metabolic CO_2 production rate	200	mL CO_2 at STPD/min
K_L	Metabolic fraction legs	0.12	
K_A	Metabolic fraction abdomen and thorax	0.65	
K_N	Metabolic fraction neck and head	0.23	
V_G	Alveolar gas volume	3	L
p_I	Partial pressure CO_2 inspired	0.3	mmHg
P_S	Standard pressure	760	mmHg
T_S	Standard temperature	273	K
T_B	Body temperature	310	K

The model has a single equilibrium point, which is easily computed. The value of the Hill function coefficient a is chosen so that the equilibrium values in compartments c_1 through c_8 and c_{15} are all equal to the typically measured arterial value of 480 mL/L. The equilibrium value of c_{14} is M/Q larger than that in the arterial compartments, giving the typical value of 520 mL/L. The remaining compartments' equilibrium values lie between these two values and are dictated by the J and K

Table 2 Average normalized blood volume shifts (mL)

	Legs	Abdomen	Thorax	Neck
Males	−145	23.5	63	58.5
Females	−365	42	270.8	52.2

fractions. This equilibrium point undergoes a supercritical Hopf bifurcation as either the gain, defined by $\mu = n/4a$, or the half-maximal ventilation/perfusion ratio, $r = \overline{\dot{V}/Q}$, is increased. When these values are low, the equilibrium point is globally asymptotically stable, representing normal and steady breathing. However, when either of these parameters is increased sufficiently, the equilibrium becomes unstable and a stable limit cycle appears, representing the cyclic breathing of CSR.

3 Blood Volume Shifts

Fluid volumes were measured in seven patients. Electrical impedance was recorded from electrodes on one side of the patient's body and converted to fluid volumes in four areas: legs, abdomen, thorax, and neck/head [8]. These measurements were taken after standing for 5 min and then again after lying down for 90 min. Blood volume shifts were computed by normalizing these measurements to 5 L of fluid for the whole body, and then taking the difference (supine minus standing) in each of the four body areas. The averaged results for males and females are shown in Table 2. Each of the 15 compartments of the model were distributed between the four measured body areas, as indicated in Table 3. The base-line volumes of Table 1 were used for the standing volumes. The computed blood volume shifts were applied proportionately to the 15 compartments and these shifted volumes were added or subtracted from the base-line volumes to yield supine volumes.

Figure 2 shows the Hopf bifurcation curves for the data averaged by gender. The axes for these plots are the two unknown parameters: The gain and the half-maximal ventilation/perfusion ratio. The former measures the sensitivity of the feedback control mechanism, while the latter is a relative measure of the average breathing rate. Below the Hopf curves, the model is at a steady equilibrium, while for parameter values above the curve the stable solution to the model is a limit cycle representing CSR. In both males and females, the supine Hopf curve lies below the standing curve, indicating that CSR is more likely to occur for patients in the supine position.

Table 3 Proportion of each compartment in the four body areas

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Legs	0	0	0	0	0.5	0	0	1	0	0	1	0	0.79	0	0
Abdomen	0	0	0	0	0.2	0	1	0	0	1	0	0	0.1	0	0
Thorax	1	1	1	0.2	0.3	0	0	0	0	0	0	0.2	0.11	1	1
Neck	0	0	0	0.8	0	1	0	0	1	0	0	0.8	0	0	0

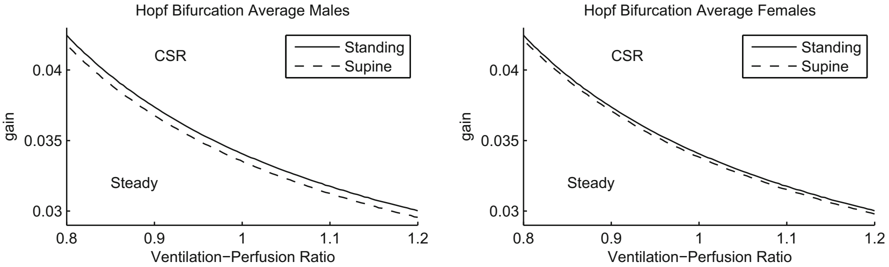


Fig. 2 Hopf bifurcation curves by gender

The difference between these curves is small but is more pronounced in males than females, which is in accord with clinical evidence. Therefore, the clinical hypothesis that fluid shifts are associated with body position can account for a higher likelihood of experiencing CSR is corroborated by the model.

Acknowledgement Fluid volumes were measured by Azadeh Yadollahi, a member of Dr. T.D. Bradley's Sleep Lab at the Toronto Rehabilitation Institute.

References

1. Atamanyk, J., Langford, W.F.: A compartmental model of Cheyne–Stokes respiration. Fields Inst. Comm. **36**, 1–16 (2003)
2. Dong, F., Langford, W.F.: Models of Cheyne–Stokes respiration with cardiovascular pathologies. J. Math. Biol. **57**(4), 497–519 (2008)
3. Goerke, J., Mines, A.H.: Cardiovascular Physiology. Raven Press, New York (1988)
4. Guyton, A.C., Hall, J.E.: Textbook of Medical Physiology, 10th edn. W.B. Saunders Company, Philadelphia (2000)
5. Levick, J.R.: An Introduction to Cardiovascular Physiology, 3rd edn. Arnold, London (2000)
6. Mohrman, D.E., Heller, L.J.: Cardiovascular Physiology, 6th edn. McGraw-Hill, New York (2006)
7. Naughton, M.T.: Pathophysiology and treatment of Cheyne–Stokes respiration. Thorax **53**(6), 514–518 (1998)
8. Redolfi, S., Yumino, D., Ruttanaumpawan, P., Yau, B., Su, M.C., Lam, J., Bradley, T.D.: Relationship between overnight rostral fluid shift and obstructive sleep apnea in nonobese men. Am. J. Respir. Crit. Care Med. **179**, 241–246 (2009)
9. Sahlin, C., Svanborg, E., Stenlund, H., Franklin, K.A.: Cheyne–Stokes respiration and supine dependency. Eur. Resp. J. **25**(5), 829–833 (2005)
10. Su, M.C., Chiu, K.L., Ruttanaumpawan, P., Shiota, S., Yumino, D., Redolfi, S., Haight, J.S., Yau, B., Lam, J., Bradley, T.D.: Difference in upper airway collapsibility during wakefulness between men and women in response to lower-body positive pressure. Clin. Sci. **116**(9), 713–720 (2009)
11. West, J.B.: Respiratory Physiology: The Essentials, 6th edn. Lippincott Williams & Wilkins, Philadelphia (2000)
12. Wilcox, M.: A model of the effects of fluid variation due to body position on Cheyne–Stokes respiration. Master's thesis, University of Guelph, Guelph, Canada (2013)

Solving a Large-Scale Thermal Radiation Problem Using an Interoperable Executive Library Framework on Petascale Supercomputers

Kwai Wong, Eduardo D'Azevedo, Zhiang Hu, Andrew Kail and Shiquan Su

Abstract We present a novel methodology to compute the transient thermal condition of a set of objects in an open space environment. The governing energy equation and the convective energy transfer are solved by the sparse iterative solvers. The average radiating energy on a set of surfaces is represented by a linear system of the radiosity equations, which is factorized by an out-of-core parallel Cholesky decomposition solver. The coupling and interplay of the direct radiosity solver using graphics processing units (GPUs) and the CPU-based sparse solver are handled by a light weight software integrator called Interoperable Executive Library (IEL). IEL manages the distribution of data and memory, coordinates communication among parallel processes, and also directs execution of the set of loosely coupled physics tasks as warranted by the thermal condition of the simulated object and its surrounding environment.

1 Introduction

The thermal content of an object sitting in an open space environment is governed by the principle of conservation of energy and its conjugating boundary conditions. As the number of thermally active surfaces of an object increases, computing the

K. Wong (✉) · A. Kail · S. Su
1122 Volunteer Blvd, University of Tennessee, Knoxville, TN 37996 USA
e-mail: kwong@utk.edu

A. Kail
e-mail: akail@utk.edu

S. Su
e-mail: ssu2@utk.edu

E. D'Azevedo
ORNL, P.O. Box 2008, Oak Ridge, TN 37831-6173 USA
e-mail: e6d@ornl.gov

Z. Hu
Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: hza8816145@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_72

energy balance related by conduction, convection, and radiation can be challenging and certainly can utilize the capability of a parallel computer. The complexity of the underlying formulations also makes it difficult for a single computer code to handle every aspect of the simulation. In this chapter, we introduce a light weight software integrator called Interoperable Executive Library (IEL) that is used to manage, coordinate, and execute the set of governing multiphysics tasks. To compute the thermal content of an object, the energy equation and the convection energy transfer are formulated by the finite element method (FEM) and solved by the sparse iterative solvers given in Trilinos [7]. The amount of radiating energy on a set of surfaces is modeled by a linear system of the radiosity equations and solved by a newly developed parallel dense matrix Cholesky decomposition procedure. The derived methodology exploits the high-throughput capability of the emergent supercomputers composed of CPUs and graphics processing units (GPUs). The radiosity matrix requires the computation of the view factors. Based on a serial view factor algorithm derived by Walton [8], we have extended the algorithms to compute the view factors on a parallel computer equipped with GPUs. The coupling and interplay of the direct radiosity solver using GPUs and the CPU-based FEM sparse solvers are handled by the IEL. The goal of the IEL is to efficiently incorporate physics solvers in a modular fashion and provides a simple application programming interface (API) for handling data transfer and scheduling. The results of a benchmark test performed on Keeneland, a GPU-based supercomputer at the National Institute for Computational Sciences (NICS)[1], is presented.

2 The Interoperable Executive Library

The IEL is a software framework used for multiphysics simulations and is designed to execute and schedule in parallel a series of physics solvers. In these multiphysics simulations, domain interaction is a common occurrence and therefore requires data and information exchange on points called shared boundaries.

Beyond its scheduling and data managing capabilities the IEL also makes use of common scientific libraries, such as the Trilinos [7], MAGMA [3], and ScaLAPACK [6] libraries. Other tools for grid generation using Cubit [1] and visualization with Paraview [5] have also been utilized. Integration of third-party solvers as modules extends the scope of the application of the library.

Figure 1 is an overview of a simulation using the IEL during runtime. A user-specific driver program initiates the execution of the simulation by passing a configuration file to the executive that starts the sequence of simulation. The IEL consists of three components: the configuration file, communicator library (COMMLIB), and executive. The configuration file defines the functionality of each simulation, the number of shared boundary conditions between different modules, and the number

[1] Web site at http://keeneland.gatech.edu/.

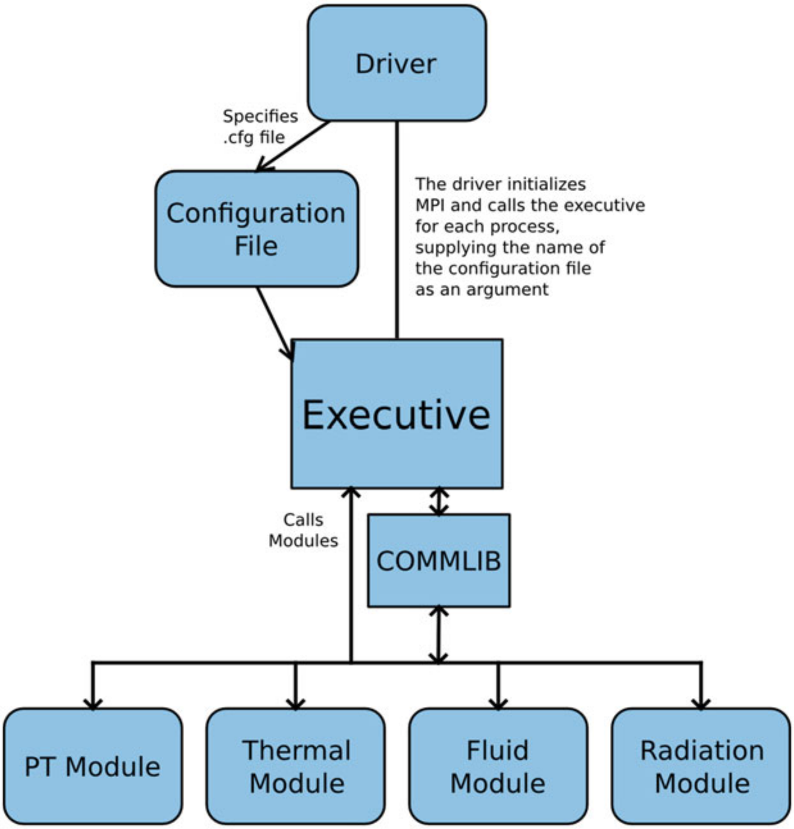

Driver
Specifies .cfg file
Configuration File
The driver initializes MPI and calls the executive for each process, supplying the name of the configuration file as an argument
Executive
Calls Modules
COMMLIB
PT Module
Thermal Module
Fluid Module
Radiation Module

Fig. 1 An overview of the Interoperable Executive Library (IEL) showing the interaction between the driver, configuration file, executive and communicator library (COMMLIB)

of processors that will be assigned for the parallel simulation. The second component is the COMMLIB that is built as a wrapper for the message passing interface (MPI) and handles the transfer of the data on the shared boundaries between modules. The third component is the executive that schedules and manages the workflow of a set of physics simulation prescribed in the configuration file.

Solving a multiphysics problem using the loosely coupled method alleviates the burden of creating a monolithic single purpose code. The IEL utilizes this method, which uses the shared boundaries as points of data exchange between any two physics solvers. Doing so allows the IEL to incorporate any independently constructed physics codes to run a multiple series of simulations, either simultaneously or in sequence, for scaling and parametric studies.

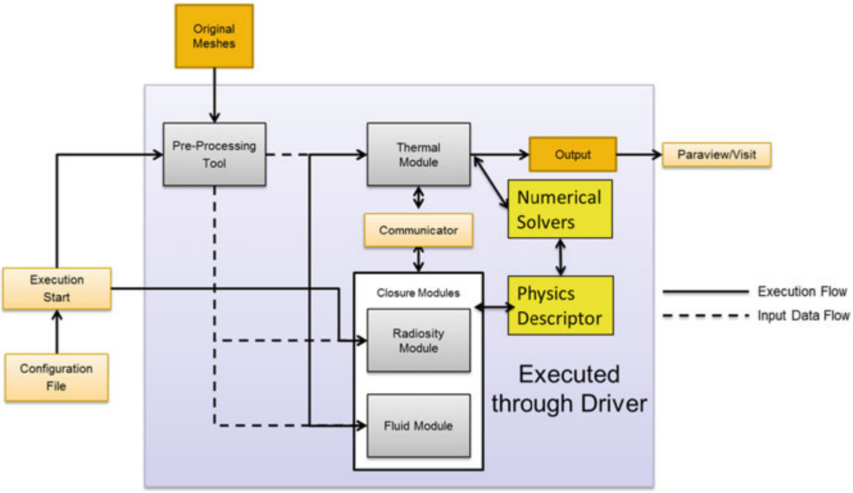


Fig. 2 A workflow diagram of the thermal fluid simulation

3 Thermal and Fluid Flow Calculations

The thermal module used in this simulation is built to simulate conjugate heat transfer on 2D surfaces in 3D space to reduce the complexity of the system and allow for generation of very complex structures such as vehicles and buildings. Figure 2 shows the workflow of the thermal fluid simulation scheduled by the IEL.

The thermal module solves the energy equations, Eqs. (1) and (2).

$$\rho C_p \frac{\partial T}{\partial t} - \nabla \cdot [k(x)\nabla T] - s(x,t) = 0, \tag{1}$$

$$k\nabla T \cdot n + h_{conv}(T - T_{atm}) + \sigma\epsilon(T^4 - T_{ref}^4) + f(t) \cdot n = 0, \tag{2}$$

where ρ, C_p, k, ϵ, and σ denote density, specific heat, conductivity, thermal emissivity, and the Stefan–Boltzmann constant, respectively. The convective terms in Eq. (2) are handled through a series of correlations that depend on the surface velocity distribution calculated by the fluid flow simulation.

The fluid module solves the potential equations, Eqs. (3) and (4), for the elemental surface velocity used by the thermal module to calculate the convective properties.

$$\nabla^2\phi = 0, \tag{3}$$

$$\frac{\partial \phi}{\partial x_i} = v_i, \quad 1 \leq i \leq n. \tag{4}$$

Both the thermal and the fluid flow simulation are formulated by the FEM and solved by sparse iterative solvers provided by the Trilinos scientific library developed at the Sandia National Laboratories.

4 Radiation Calculations

Assuming a gray diffuse environment, the amount of average radiating energy to and from an object can be represented by a system of radiosity equations, Eqs. (5) and (6).

$$\epsilon_i R_i A_i = E_i A_i + \rho_j \sum_{j=1}^{n} R_j A_j F_{ij}, \quad \text{for each } i = 1, 2, \ldots, N, \tag{5}$$

$$E_i = \epsilon \sigma T_i^4, \tag{6}$$

where i is the surface which the radiosity is being calculated, j represents a single surface, E is the emitted energy, ρ is the the reflectivity, ϵ is the emissivity and equals to $(1 - \rho)$, and F_{ij} is the view factor between surface i and j.

Equation (5) can be rewritten as follows:

$$\mathbf{G}x = (\delta_{ij} A_i - \phi_i A_i F_{ij}) R_i = \sigma T_i^4 = b, \tag{7}$$

which is a system of linear equation, $\mathbf{G}x = b$, in which $\mathbf{G}$ is a symmetric positive definite matrix because of the reciprocity property of the view factors, $A_i F_{ij} = A_j F_{ji}$.

The radiosity matrix requires the computation of the view factors, which depends only on the geometry and orientation of the two interacting surfaces. Walton [8] has listed several commonly used algorithms to compute the view factors in his View3D code. Calculating the view factors is a compute intensive process but can be done in parallel. Based on the serial view factor algorithms listed by Watson, we have extended the View3D code to compute the view factors on a parallel computer equipped with GPU accelerators. View3D applies modified area integration methods to calculate the partially obstructed surfaces. Every processor calculates the portion of view factors arranged in the distributed 2D block cyclic data decomposition used in ScaLAPACK. The Cholesky factorization is then used to solve the system of the radiosity equations. A GPU-based out-of-core scheme using cuBLAS [4] and ScaLAPACK library has been developed to perform the Cholesky factorization of matrix $\mathbf{G}$.

Out-of-core factorization methods have been studied in the past [2]. The method uses the secondary memory, in here the host memory in the CPU node, to augment the storage and solving of a large matrix which exceeds the limit of the primary memory, here the memory of a GPU. In this chapter, we adapt the left-looking out-of-core algorithm for the Cholesky factorization that seeks to minimize the data transfer between the CPU host and the GPU device memory. Central to this algorithm is also an in-core parallel factorization method that operates primarily on the GPU with minimal communication between GPUs. The primary variant is that we assume that the portion of the matrix $\mathbf{G}$ belonging to a CPU processor is too large to be fully held entirely in GPU device memory. Thus, some data movement of the matrix between the CPU and the GPU will be necessary, but must be minimized to achieve good performance.

Table 1 Speed up of L-shape test on kraken

Nnumber of processors (M)	Time(s) for potential obstruction detecting +	Time(s) for view factor computing
	View factor computing	
4	140.95	113.22
8	101.59	56.85
16	74.99	29.34
32	60.41	14.42
64	53.24	7.14
128	49.80	3.69
256	48.13	1.78
1024	47.25	0.44

Table 2 Performance of Cholesky factorization using 12 cores per node with 12 MPI tasks per node and block size $NB = 128$

Processor grid	Matrix size N	Performance per GPU (unit: GFlops/s)
6×6	57,600	165
12×12	116,736	145
24×24	216,576	140

5 Results and Discussions

We used a simple L-shape benchmark test to examine the speed up of the parallelized View3D code. There are 20,000 surfaces in the L-shape plate. Table 1 shows the speed up is almost linear in computing the view factors. However, the time to construct the list of obstruction is almost constant for the L-shape. Because there is no obstruction in the L-shape test, the algorithm requires $O(N^2)$ instructions to complete the check list. Table 2 shows the performance of the parallel out-of-core Cholesky factorization reaches 160 Gflops/s per GPU on Keeneland.

Acknowledgment This research project was supported by the National Science Foundation and the Department of Energy under Contract No. DE-AC05-00OR22725.

References

1. Cubit. https://cubit.sandia.gov/
2. D’Azevedo, E., Hill, J.: Parallel LU factorization on GPU cluster. Proc . Comput. Sci. **9,** 67–75 (2012)
3. Dongarra, J.: MAGMA User’s Guide. Innovative Computing Laboratory (2009)
4. Nvidia: https://developer.nvidia.com/cublas

5. Paraview. www.paraview.org
6. Scalapack. http://netlib.org/scalapack/slug/index.html
7. Trilinos: http://trilinos.sandia.gov
8. Walton, G.N.: Calculation of obstructed view factors by adaptive integration. Tech. rep., National Institute of Standards and Technology (2002)

Optimal Transport and Placental Function

Qinglan Xia, Carolyn Salafia and Simon Morgan

Abstract The human newborn is a reflection of the entirety of nutrients transferred from the maternal to the fetal circulation across the placenta during gestation. By extension, birth weight and newborn health depend on placental function. The goal of this chapter is to introduce the use of optimal transport modeling to study the expected effects of (i) placental size, (ii) placental shape (separate from size), and (iii) the position of insertion of the umbilical cord, on birth weight and placental functional efficiency. For each placenta (N = 1110), a total transport cost based on all measurements (i), (ii), and (iii) is given by the model. This computed cost is highly correlated with measured birth weight, placenta weight, the fetal–placental weight ratio (FPR), and the metabolic scaling factor beta. Next, a shape factor is calculated in a model of the total transport cost if each placenta were rescaled to have a unit area chorionic plate (thus separating shape from size). This shape factor is also highly correlated with birth weight, and after adjustment for placental weight, is highly correlated with the metabolic scaling factor beta.

1 Introduction

The human newborn is the reflection of the sum total of oxygen and nutrients transferred from the maternal to the fetal circulation across the placenta during gestation. By extension, birth weight depends on placental function. The goal of this chapter is to apply optimal transport modeling to quantify effects of (i) placental size, (ii)

Q. Xia (✉)
Department of Mathematics, University of California at Davis, Davis, CA 95616 USA
e-mail: qlxia@math.ucdavis.edu

C. Salafia
Placental Analytics, LLC, Larchmont NY 10538, USA
Institute for Basic Research, Staten Island, NY, USA
e-mail: carolyn.salafia@gmail.com

S. Morgan
Los Alamos National Laboratory, Los Alamos, NM 87544, USA
e-mail: morga084@gmail.com

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_73

placental shape, and (iii) the position of insertion of the umbilical cord on the chorionic disk surface, on birth weight. This size, shape, and position data was readily available from measurements from photographs of 1110 placentas from a University of North Carolina birth cohort collected in the middle of the last decade, which has been extensively studied in e.g., [3, 11] and references therein.

The measures, (i),(ii), and (iii) above have expected effects on the energy required to pump blood across the placenta. Generally in any transport, and we assume also in the placenta, the less distance the blood has to travel, the less energy needs to be expended to pump it. Therefore, the predicted optimum shape for the chorionic plate to minimize transportation energy is a circle with a centrally inserted umbilical cord. If the umbilical cord insertion point is eccentric within a circular chorionic plate, then overall the blood will have farther to travel to and from to the umbilical cord. Also if the chorionic plate is not circular, but elliptical or lobated, then again, overall, the blood will have farther to travel and so more energy expenditure will be needed. Thus, one may expect that placental shape and location of umbilical cord are important factors in determining the energy needed to pump blood across the fetal–placental circulation. From this, one would also assume that given a larger placenta, more blood would be transported over a longer distance, with more energy required for pumping.

In this chapter we simulate a vascular tree structure for each placenta, in a simplified form by an idealized optimal transport network. For this network there is an associated total transport cost C computed by the model. This cost C represents the total work done by the heart of the fetus to pump blood across the placenta. We find a high correlation between C and measured birth weight, placenta weight, the fetal–placental weight ratio (FPR) and the metabolic scaling factor beta. Also, a shape factor S is computed by the model which would be the total transport cost if a placenta was rescaled to have a unit area chorionic plate. This shape factor S is also highly correlated with birth weight, and after adjustment for placental weight, is highly correlated with the metabolic scaling factor beta.

2 Modeling Method

The optimal transportation problem aims at finding an optimal way to transport materials from the source to the target. An optimal transport path introduced in [7] is a mathematical concept used to model tree-shaped branching transport networks. Transport networks with branching structures are observable not only in nature as in trees, blood vessels, river channel networks, lightning, etc. but also in efficiently designed transport systems such as used in railway configurations and postage delivery networks. Recently, mathematicians (e.g., [1, 2, 4, 7]) have shown great interest in modeling these transport networks with branching structures. Applications of optimal transport paths may be found in [8] and [9]. A related interesting approach is given in [5] which investigates thermodynamic properties of optimal transport networks while [6] investigates thermodynamic properties of measured human placenta major blood vessel networks. In this chapter, we will model the blood vessel structure of a placenta via an optimal transport path.

As stated in [11], 1110 placentas were collected by an academic health center in central North Carolina. For each placenta, a trained observer captured a series of x,y coordinates that marked the site of the umbilical cord insertion and the perimeter of the fetal surface. To simulate vascular structures for the placentas, we apply the modeling method of ramified optimal transportation to each placenta.

An idealized transport network, which simulates an optimal vascular structure for that placenta, is computed based on the measurements of the placenta. This branched network provides a means of transporting blood between the whole chorionic plate surface and the umbilical cord. This single network for a placenta may be viewed as a representation of either an optimal vein network or, by reversing directions of flow, an optimal arterial network. In the absence of more detailed information about blood supply, we assume a uniform supply of blood per unit area over the whole surface of the placenta. We also model the placenta by a region in the plane because the data is from photographs of the placenta flat on a table, rather than in the curved inside surface of the uterus. The idealized transport network is a branched network of straight segments e_i each with a capacity weighting w_i and a direction of flow. For each branch point, the sum of flows in must equal the sum of flows out. Since there are many ways to construct a transport network, we need to find an optimal network which minimizes the amount of work done in pumping blood through the network. In the model of ramified optimal transportation, we use the cost function $(w_i)^\alpha l_i$ for each edge e_i of length l_i where α is a branching parameter $(0 \leq \alpha < 1)$. Technically, as x^α is strictly concave for this range of α, this ensures that branched structures will emerge and corresponds to the general principle of favoring transportation in groups and branched vessel structures. The total cost for each transport network, which reflects the work done to pump the blood, is the sum of the costs for each edge. Using algorithms stated in [10], for any fixed α we can build an approximating optimal transport path for a placenta using its measurements (e.g., Fig. 1, left with $\alpha = 0.85$). Then we may calculate the associated total transport cost

$$C = \sum (w_i)^\alpha l_i$$

for that placenta.

For the calculations, we chose the value of $\alpha = 0.85$ so that, for a round placenta with a centrally inserted umbilical cord, six branches will emerge from the umbilical cord. This is consistent with the typical observation that four to six branches emerge from the umbilical cords in normal round placentas. We also used the uniformly distributed point sources as shown on the left of Fig. 1. This choice was made because if random distributions of point sources were used, the model would give different values of total transport cost C for the same placenta each time the model was run. The total transport cost C for each placenta depends upon shape, size, and umbilical cord position. We want to investigate the effect of the shape and umbilical cord position independently from size. To do it, we consider

$$S = \frac{C}{A^{0.5+\alpha}},$$

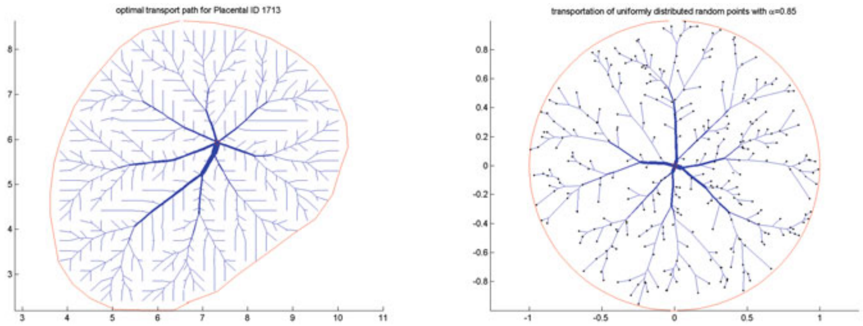

optimal transport path for Placental ID 1713
8
7
6
5
4
3
3
4
5
6
7
8
9
10
11
transportation of uniformly distributed random points with α=0.85
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.5
0
0.5
1

Fig. 1 Examples of modeling blood vessels of a placenta by a nearly optimal transport network from the placenta surface to the umbilical cord. The distribution of the blood source over the surface of the placenta is uniform over lattice points of a fine regular grid in the example on the *left*, and is on randomly placed points in the example on the *right*

where A is the area of the placenta. Note that, the value of S is a function of shape and cord position, and is independent of size. Indeed, suppose D_1 and D_2 are two placentas of the same shape. Then D_1 can be viewed as a rescale of D_2 with a length scaling factor $\lambda > 0$. Thus, $Area\,(D_1) = \lambda^2 Area\,(D_2)$. Let G_1 and G_2 be the corresponding optimal transport networks for D_1 and D_2. One may also show that the total cost C_1 for G_1 is the total cost C_2 for G_2 multiplied by $\lambda^{2\alpha+1}$. As a result,

$$S_1 = \frac{C_1}{A_1^{0.5+\alpha}} = \frac{\lambda^{2\alpha+1} C_2}{\left(\lambda^2 A_2\right)^{0.5+\alpha}} = \frac{C_2}{A_2^{0.5+\alpha}} = S_2.$$

We call S the *shape factor* of the placenta. As a result, the total transport cost C can be expressed as the product of two independent variables: $C = S * A^{0.5+\alpha}$.

3 Results

For each of the 1110 placentas, the associated birth weight B of the fetus and placental weight P are also available. We applied the above method to calculate the total transport cost C and the shape factor S. Placental functional efficiency is typically measured either by the FPR $= \frac{B}{P}$ or by the metabolic scaling factor beta, $\beta = \frac{\ln B}{\ln P}$.

As shown in Table 1, total transport cost is highly correlated with birth weight, placental weight, FPR, and beta. Total transport cost C is positively correlated with birth weight as expected given that C primarily reflects placental size, and on average will vary with larger and smaller placental and fetal weights.

On the other hand, the shape factor S is negatively correlated with birth weight as we would expect consistent with our hypothesis that a high S (and therefore an irregular shape with greater deviations of cord location and/or irregularities of perimeter) significantly impairs placental efficiency for nutrient transportation under

Table 1 Pearson's correlations

		Birth weight	Placental weight	FPR	Beta
Total transport cost C	Pearson correlation	0.421	0.489	−0.154	0.272
	Sig. (2 tailed)	0.000	−0.000	0.000	0.000
Shape factor S	Pearson Correlation	−0.080	−0.020	−0.056	0.039
	Sig. (2 tailed)	0.008	0.508	0.062	0.192

Table 2 Regression coefficients (point estimate of effect) for total transport cost and shape factor on birth weight (Model 1) and after adjustment for placental weight (Model 2)

		Unstandardized coefficients			
Model		Birth weight	Std. error	t	Sig.
1	(Constant)	2483.951	54.964	45.193	0.000
	Total transport cost C	0.590	0.038	15.400	0.000
2	(Constant)	1546.922	64.502	23.983	0.000
	Total transport cost C	0.210	0.037	5.639	0.000
	Placental weight	3.307	0.158	20.958	0.000
1	(Constant)	3693.731	152.020	24.298	0.000
	Shape factor	−594.053	222.689	−2.668	0.008
2	(Constant)	1985.163	134.301	14.781	0.000
	Shape factor	−501.411	173.258	−2.894	0.004
	Placental weight	3.734	0.139	26.837	0.000

the conditions of an optimal transport network. In this sample the effect of shape factor S on birth weight is not paralleled by a correlation of abnormal shape with placental weight, with only trends to correlations with FPR and beta.

After adjustment for placental weight in regression analysis, the significant relationships of both total transport cost and the shape factor on birth weight remained (see Table 2). Both variables were also highly correlated with the metabolic scaling factor beta after adjustment for placental weight (see Table 3).

For total transport cost, we do not expect model 2 to be greatly better than model 1, since placental area is factored into total transport cost and thus total transport cost in isolation includes placental size. However, the shape factor S does not reflect placental size. Therefore, we do expect the inclusion of placental weight into model 2 to make a large difference as compared with model 1. The shape factor does not factor in placental area, and so does not reflect placental size. Model 1 includes shape factor S only, and thus no influence of placental size. Model 2 (which includes placental weight as a covariate) does. In both models shape factor S has a significant point estimate of effect on birth weight. The second model has a somewhat reduced point estimate of effect for shape factor S, with a smaller standard error, making this slightly smaller estimate of effect more precise.

Table 3 Regression coefficients (point estimate of effect) for total transport cost and shape factor on beta (Model 1) and after adjustment for placental weight (Model 2)

Model		Unstandardized coefficients		t	Sig.
		Beta	Std. error		
1	(Constant)	0.731	0.002	334.290	0.000
	Total transport cost	1.43E-005	0.000	9.374	0.000
2	(Constant)	0.680	0.002	324.002	0.000
	Total transport cost	$-6.2E-006$	0.000	−5.104	0.000
	Placental weight	0.000	0.000	34.606	0.000
1	(Constant)	0.743	0.006	130.033	0.000
	Shape factor	0.011	0.008	1.306	0.192
2	(Constant)	0.667	0.004	152.921	0.000
	Shape factor	0.015	0.006	2.670	0.008
	Placental weight	0.000	0.000	36.544	0.000

Table 3 shows the same models of total transport cost and shape factor S predicting beta. Total transport cost is correlated with beta (placental functional efficiency). Model 2 includes placental weight; the distribution of beta varies with placental weight (heteroscedastic). Therefore, even though beta is calculated from placental weight, it is reasonable to include placental weight as a covariate. The point estimate of effect is reduced after adjustment for placental weight but remains highly statistically significant. Shape factor S is uncorrelated with beta in univariate regression, consistent with the results of correlation. Placental surface shape in isolation, out of context of other parameters of the placenta, would hardly be expected to be a predictor of placental functional efficiency. However, the more regular the shape for a given placental weight (Model 2) the less the beta, and the larger the placenta relative to the birth weight (reflecting poorer functional efficiency). Thus, while shape does not have independent effects on beta, the rounder any placenta is (the lower the shape factor S) at a given weight, the more efficient the placenta.

Acknowledgement The work of Qinglan Xia is supported by the NSF grant DMS-1109663. The work by Simon Morgan was funded by the Department of Energy through the LANL/LDRD office #X1LJ, LA-UR 10-00739. Thanks to IPAM at UCLA for additional support.

References

1. Bernot, M., Caselles, V., Morel, J.M.: Optimal Transportation Networks: Models and Theory, Lecture Notes in Mathematics, Springer, Vol. 1955 (2009)
2. Brancolini, A., Buttazzo, G., Santambrogio, F.: Path functions over Wasserstein spaces. J. Eur. Math. Soc. **8**(3), 415–434 (2006)
3. Gill, J.S., Salafia, C.M., Grebenkov, D., Vvedensky, D.D.: Modeling oxygen transport in human placental terminal villi. J. Theor. Biol. **291**, 33–41 (2011 Dec 21)

4. Maddalena, F., Solimini, S., Morel, J.M.: A variational model of irrigation patterns. Interface. Free Bound. **5**(4), 391–416 (2003)
5. Seong, R.K., Salafia, C.M., Vvedensky, D.D.: Statistical topology of radial networks: a case study of tree leaves. Philosophical Magazine, iFirst, 1–16 (2011)
6. Seong, R.K., Getreuer, P., Li, Y., Girardi, T., Salafia, C.M., Vvedensky, D.D.: Statistical geometry and topology of the human placenta. Advances in Applied Mathematics, Modeling, and Computational Science, Fields Institute Communications, vol. 66, pp. 187–208 (2013)
7. Xia, Q.: Optimal paths related to transport problems. Commun. Contemp. Math. **5**(2), 251–279 (2003)
8. Xia, Q.: The formation of tree leaf. ESAIM Control Optim. Calc. Var. **13**(2), 359–377 (2007)
9. Xia, Q., Unger, D.: Diffusion-limited aggregation driven by optimal transportation. Fractals **18**(2), 247–253 (2010)
10. Xia, Q.: Numerical simulation of optimal transport paths. arXiv:0807.3723. The Second International Conference on Computer Modeling and Simulation. Vol. 1, 521–525, 2010
11. Yampolsky, M., Salafia, C.M., Shlakhter, O.: Probability distributions of placental morphological measurements and origins of variability of placental shapes. Placenta **34**(6), 493–496 (2013)

Localized Band-Limited Representation and Robust Interpolative Image Manipulation

H. Xiao, M. C. Gonzalez and N. Fugate

Abstract In this chapter, we describe an image representation framework based on which a robust, nonparametric interpolation method for filling in "missing" information of an image can be performed. As in an earlier work, this approach utilizes a class of localized band-limited functions that are compact in both image and frequency domains. However, the current algorithm may be carried without a statistical classifier such as a K-means algorithm, which was employed in our previous work. After a brief description of our approach, results are given to show its efficacy in a few use cases.

1 Introduction

With the rapid development of imaging devices and ever increasing computational power, a wide variety of interesting image analysis problems have arisen in fields ranging from medicine, chemistry, geophysics, satellite imagery, and remote sensing to digital photography. Whether for the purpose of segmenting cancerous cells from healthy ones via hyperspectral imaging, or reconstructing a three-dimensional model of proteins from two-dimensional slices, features of objects in an image often need to be modeled, detected, extracted, enlarged, and cataloged. The effectiveness of the approaches generally depends upon the underlying mathematical model used for representing the images.

H. Xiao(✉)
Department of Computer Science, University of California,
One Shields Ave, Davis, CA 95616, USA
e-mail: hxiao@ucdavis.edu

M. C. Gonzalez
Department of Electrical and Computer Engineering, University of California,
One Shields Ave, Davis, CA 95616, USA
e-mail: margonzalez@ucdavis.edu

N. Fugate
University of California, One Shields Ave, Davis, CA 95616, USA
e-mail: nfugate@ucdavis.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_74

Although images are seldom truly band-limited (i.e., have compactly supported Fourier transforms), they are generally considered "piece-wise smooth" and are frequently modeled *locally* as such by polynomials and trigonometric functions. In particular, image representation and band-limited functions are no strangers to one another. Indeed, discrete cosine transform (DCT) is part of the popular image format JPEG; many other algorithms in image processing (filtering, encoding, edge detection, texture analysis, etc.) use Fourier analysis as a basic tool. Since image features such as edges, corners, shadows, etc., are limited in the spatial domain, the ideal representation should also be *localized* in spatial domains. In other words, the ideal bases for image representation should be simultaneously compact in both image and frequency domains.

In this chapter, we present a mathematical framework for representing images in localized band-limited functions, in particular, with prolate spheroidal wave functions (PSWFs). We first present relevant mathematical facts of the PSWFs in Sect. 2, and introduce our representation framework and the interpolation algorithm in Sect. 3. Results of using this approach without the combination of a statistical classifier for filling in missing data of an image are shown in Sect. 4. Finally, we give conclusions in Sect. 5.

2 A Localized Band-Limited Basis

In this section, we introduce a basis of band-limited functions, the PSWFs. We summarize relevant facts of PSWFs below; for more detailed discussion about these functions, see, for example, [1–3, 6].

PSWFs ψ_n^c are eigenfunctions of the finite Fourier integral operator $F_c : L^2[-\Omega, \Omega] \to L^2[-T/2, T/2]$ defined by the formula

$$F_c(\varphi)(t) = \int_{-\Omega}^{\Omega} e^{icts}\, \varphi(s)\, ds \tag{1}$$

with $c = \Omega\, T/2$. To be more specific, for any integer $n \geq 0$, there exists a complex number λ_n^c and a corresponding real valued function ψ_n^c such that

$$\lambda_n^c\, \psi_n^c(t) = \int_{-\Omega}^{\Omega} e^{icts}\, \psi_n^c(s)\, ds \tag{2}$$

for all $t \in [-T/2, T/2]$. Simple algebraic manipulations show that ψ_n^c are also eigenfunctions of the "sinc" integral operator $Q_c : L^2[-\Omega, \Omega] \to L^2[-T/2, T/2]$ given by the formula

$$Q_c(\varphi)(t) = \frac{1}{\pi} \int_{-\Omega}^{\Omega} \frac{\sin c\,(t-s)}{t-s}\, \varphi(s)\, ds. \tag{3}$$

In other words,

$$\mu_n^c\, \psi_n^c(t) = \frac{1}{\pi} \int_{-\Omega}^{\Omega} \frac{\sin c\,(t-s)}{t-s}\, \psi_n^c(s)\, ds \tag{4}$$

where μ_n^c are real positive numbers (see, for example [4]). For all $c > 0$ and all integer n, $0 < \mu_n^c < 1$. Ordering μ_n^c to be strictly decreasing such that $\mu_0^c > \mu_1^c > \ldots > \mu_n^c \ldots$, we denote by $\psi_0^c, \psi_1^c, \ldots, \psi_n^c, \ldots$ the corresponding functions, and call ψ_n^c the n-th order PSWFs.

It is well known that the PSWFs form an orthonormal basis for band-limited functions on the real line. In addition, $\{\psi_n^c\}$ form a basis for all functions that are square integrable on the interval $[-T/2, T/2]$. That is, for any function ϕ defined on the interval $[-T/2, T/2]$ such that

$$\int_{-T/2}^{T/2} |\phi(t)|^2 dt < \infty,$$

we have

$$\phi(t) = \sum_{n=0}^{\infty} \alpha_n \, \psi_n^c(t), \tag{5}$$

for any $c > 0$. We call (5) the Prolate series of ϕ.

Suppose that ϕ is a "smooth" function. It has been shown that the expansion coefficient α_n is proportional to μ_n in (3) for sufficiently large c, and μ_n decays exponentially for n being sufficiently large (see [6]). The prolate series of ϕ can then be truncated to any desirable precision.

3 Image Representation and Interpolation with Localized Band-Limited Functions

The theory of band-limited functions based on PSWFs in one-dimension is completely generalizable to domains in higher dimensions. This theory is briefly summarized in this section. Due to space limitation, we omit all proofs.

For image representation, we are interested in rectangular regions, which are Cartesian separable. One can construct a basis on such domains using the tensor product rule. For example, suppose that $\{\psi_m^c(x), m = 1, 2, \ldots,\}$ and $\{\psi_n^c(y), n = 1, 2, \ldots,\}$ are each a PSWF basis defined on the interval $[-1, 1]$ with band-limited c. Then the two dimensional functions $\phi_{m,n}$ defined as

$$\phi_{m,n}(x, y) = \psi_m^c(x)\psi_n^c(y)$$

for all $(x, y) \in [-1, 1] \times [-1, 1]$ and all natural m's and n's are the tensor products of $\{\psi_m^c(x)\}$ and $\{\psi_n^c(y)\}$. In Fig. 1, we show a collection of such functions.

It can be easily shown that $\{\phi_{m,n}(x, y)\}$ form a basis for all functions that are square-integrable on $[-1, 1] \times [-1, 1]$. In addition, their Fourier transforms are compactly supported. Therefore, they are also a basis for band-limited functions. Parallel to the situation in one-dimension, the prolate series

$$f(x, y) = \sum_{m,n=0}^{\infty} \alpha_{m,n} \, \phi_{m,n}(x, y), \tag{6}$$

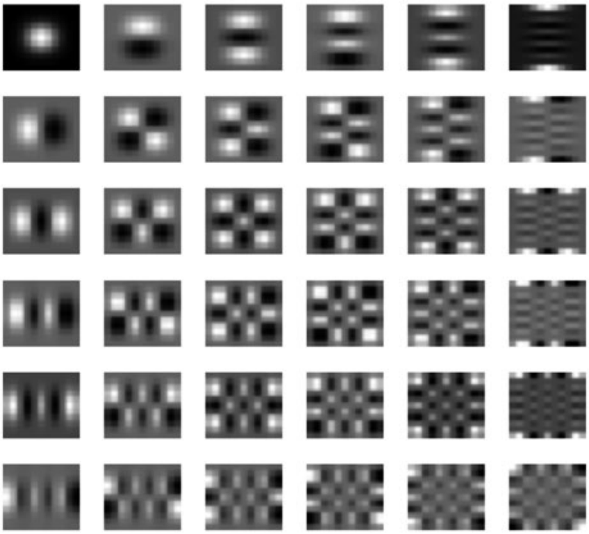


Fig. 1 A part of a Cartesian separable bandlimited basis. Here, $c = 9$, $0 \leq m, n \leq 5$. The basis functions are evaluated on a 11×11 grid

of the two-dimensional function f may be truncated for sufficiently large m and n.

Considering an intensity image as the evaluations of a continuous function $\bar{I}(x, y)$ at discrete locations $x = k$ and $y = l$, where $\bar{I}(k, l)$ are the pixel values of the image at horizontal and vertical pixel indices k and l, the image value at an arbitrary location (k', l') may be evaluated (or interpolated), as soon as the truncated expansion (6) is found. This interpolation is stable in this representation framework, as long as the basis functions "well represent" the given image. In the event that some of the image pixels are "corrupted" or otherwise become difficult to represent with the said basis, we should expect jumps between the image and the "interpolated image", where there is "missing information." Therefore, we can use a thresholding method to "predict" which pixels have likely been corrupted. We will then use a second pass of the interpolation algorithm without the "corrupted" pixels to finally interpolate our image. Using this procedure, no statistical classification methods or explicit masks for identifying the locations of the missing data are needed, which were required in our previous work [5]. As before, we process the image in nonoverlapping blocks, as no Gibbs phenomenon will be present, and the procedure is still nonparametric.

4 Results

We implemented our two phase interpolation algorithm described above. To illustrate the effectiveness of the approach, we show in Fig. 2 an example of the algorithm "denoising" an image of the eye. We introduced an additive Gaussian noise (with the average μ relative to the average pixel intensity and σ being 0.2μ) to about 10 % of the pixels of the image. Our algorithm correctly identified 9.080 % of the pixels

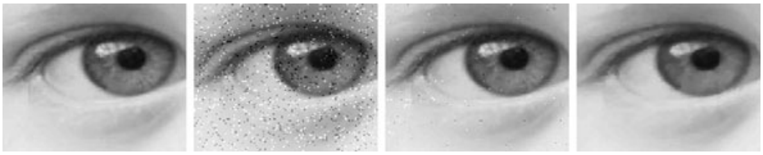


Fig. 2 From *left* to *right*: **a** the original image, **b** the corrupted image of the eye with 10 % noise, **c** the interpolated image of the eye with $c = 6$, and **d** the reconstructed image of the eye with a median filter

as being corrupted when 9.718 % of the pixels were actually corrupt. Our algorithm also identified 0.797 % of the pixels as corrupted when they were not. In comparison, we also implemented a simple median deblur-filter, which alters every pixel indiscriminately. The reconstructed image of our algorithm is noticeably sharper than that of the deblur-filter, especially in areas with more textures (e.g., near the iris). We further tested our method on additional cases of the Gaussian additive noise with several μ's and σ's. We list the results in Table 1. For each case, the additive Gaussian noise was created with the mean (relative to the average image intensity value) and standard deviation (relative to the mean) given by μ and σ, respectively. The mean squared errors (MSEs) and peak signal-to-noise ratios (PSNRs) are reported for the corrupted images and the revised image in columns 5 and 11, and columns 6 and 12, respectively. Throughout the experiment, we set $c = 7$ and set blocksize to 13. The % Noise values describe the targeted percentages of the pixel locations that were altered by the corruption process. As can be seen, the reconstruction from the noisy images is robust, and is satisfactory in general regardless of the local texture around the missing pixels. In Fig. 3, we show a result of the algorithm for "in-painting" the missing region of the image of a girl, with the missing pixel location not given.

5 Conclusions

We describe an approach for representing images in band-limited functions that are also spatially concentrated. Using this approach, we develop an interpolation procedure with two phases, which can be used in supplying missing (or corrupted) image data, without prior knowledge or explicit statistical classifications of missing pixel locations. We show that, in a range of noisy environments, the algorithm performs effectively. Therefore, in the event that noise does not overlap significantly with the original image in its frequency profile, our method is simple and effective.

Acknowledgments The authors of this work were in part supported by NSF DMS-1016712.

Table 1 Mean squared error (MSE) and peak signal-to-noise ratio (PSNR) of corrupted, and revised images using our method, and a deblurring method for several parameter settings

μ	σ	*%Noise*	*Type*	*MSE*	*PSNR*	μ	σ	*%Noise*	*Type*	*MSE*	*PSNR*
1.0	0.2	05	Corrupted Revised	0.014 0.001	25.378 36.506	0.5	0.2	05	Corrupted Revised	0.004 0.001	29.318 35.233
		10	Corrupted Revised	0.025 0.003	22.488 33.867			10	Corrupted Revised	0.009 0.025	26.189 32.098
		15	Corrupted Revised	0.038 0.006	20.538 30.163			15	Corrupted Revised	0.0143 0.0035	24.478 30.553
		20	Corrupted Revised	0.053 0.010	19.412 27.494			20	Corrupted Revised	0.0194 0.0054	23.150 28.704
1.0	0.3	05	Corrupted Revised	0.0164 0.001	24.335 34.495	4.0	0.1	05	Corrupted Revised	0.162 0.0002	13.926 43.713
		10	Corrupted Revised	0.029 0.004	21.291 29.525			10	Corrupted Revised	0.359 0.001	10.466 33.987
		15	Corrupted Revised	0.048 0.009	19.307 26.519			15	Corrupted Revised	0.521 0.008	8.845 26.770
		20	Corrupted Revised	0.0617 0.0133	18.217 24.859			20	Corrupted Revised	0.700 0.025	7.569 21.971

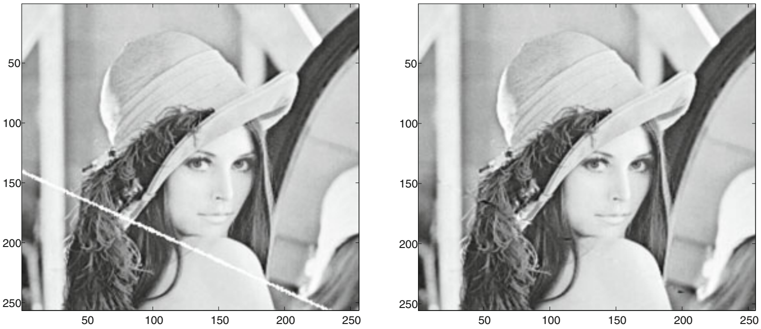


Fig. 3 Reconstruction of Lena (256×256) after a line scratch. *Left* The image with a scratch of about 3-pixel in width. *Right* The reconstructed image with $c = 6$, and M, N both being 7. The blocksize is 15

References

1. Landau, H.J., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis, and uncertainty - ii. Bell Syst. Tech. J. **40**:65–84 (1961)
2. Landau, H.J., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis, and uncertainty - iii: the dimension of space of essentially time and band-limited signals. Bell Syst. Tech. J. **41**:1295–1336 (1962)
3. Osipov, A., Rokhlin, V., Xiao, H.: Prolate Spheroidal Wave Functions of Order Zero: Mathematical Tools for Bandlimited Approximation. Springer, New York (2013)

4. Slepian, D., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis, and uncertainty - i. Bell Syst. Tech. J. **40**:43–63 (1961)
5. Wilson, R., Xiao, H.: Image interpolation based on the prolate spheroidal sequences. Proc. 7th IMA Conference on Mathematics in Signal Processing, Cirencester, United Kingdom (2006)
6. Xiao, H., Rokhlin, V., Yarvin, N.: Prolate spheroidal wave functions, quadrature and interpolation. Inverse Probl. **17**:805–838 (2001)

A Monte Carlo Measure to Improve Fairness in Equity Analyst Evaluation

John Robert Yaros and Tomasz Imieliński

Abstract The *Wall Street Journal*'s "Best on the Street," StarMine and many other systems measure analyst stock-rating performance using variations on a method we term the "portfolio method," whereby a synthetic portfolio is formed to track the analyst's ratings. At the end of the evaluation period, analysts are compared by their respective portfolio returns. Of the pitfalls to this method, one most troubling is that the analysts are generally covering different sets of stocks over different time periods. Thus, each analyst has access to different opportunities and just comparing portfolio values is unfair. In response, we present a Monte Carlo (MC) method where, for each analyst, we generate numerous "pseudo-analysts" with the same coverage over the same time periods as the real analyst. Using this method, we are better able to compare analysts, adjusted for their individual opportunities. We draw comparisons between our results and the results from existing systems, showing that those systems are less precise in reflecting analyst performance.

1 Introduction

Numerous systems for evaluating stock analysts have emerged over the years. This reflects the investor's desire to know which analysts are the best predictors of future stock behavior. Good predictions can mean the investor can achieve higher returns, so s/he is willing to pay substantially for such advice as long as s/he perceives it to be the most accurate. At the same time, measuring analyst performance is not straightforward. Each analyst likely covers a subset of stocks, such as major pharmaceutical companies, and the subsets of stocks are nearly always different across research firms employing different analysts. For example, one "retail" analyst may cover Walmart, Target, and Costco, while a retail analyst at another firm covers Walmart, Best Buy, and RadioShack. Moreover, the stocks covered by each analyst

J. R. Yaros (✉) · T. Imieliński
Department of Computer Science, Rutgers University, New Brunswick, NJ, USA
e-mail: yaros@cs.rutgers.edu

T. Imieliński
e-mail: imielins@cs.rutgers.edu

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_75

may vary in time, so making comparisons for a specified interval can be difficult since the composition of stocks covered can change frequently throughout the interval.

The de facto approach[1] to handling these challenges has been the “portfolio method,” where a synthetic portfolio is created to track ratings made by the analyst. For example, when the analyst gives a positive rating, the portfolio goes long one unit of that stock. For negative ratings, a short unit is added to the portfolio. These positions are exited when the rating ends, such as when the analyst stops coverage. The intent is that the portfolio value will reflect the accuracy of the analyst’s decisions. At the end of a given time period, analysts can be ranked with the belief that the most accurate analyst will have the highest portfolio value.

We find three inter-related shortcomings. First, each analyst covers different stocks and, thus, has access to different opportunities. So, the portfolio return for one analyst may be higher than another analyst simply because his/her stocks have greater price changes. Second, while positive and negative ratings have counterpart actions of buying and selling in the portfolio, neutral ratings do not have a clear action. A frequent approach is to simply ignore them. Another approach is to invest in a benchmark asset such that returns of the overall portfolio are diluted. It is true that returns may be lowered by missed opportunity, but again considering that not all analysts cover the same stocks, the missed opportunity may not be reflected when comparing to other analysts if few or no other analysts covered that stock. Third, to interpret portfolio return, one must have reference to the movement of the underlying stocks. For example, suppose an analyst covering only one stock has a portfolio return of 5 %. This might be excellent if perfect predictions would lead to a 6 % return overall during the period but would be much weaker if 60 % was possible.

In recognition of these issues, we present a MC approach that harks back to a 1933 study by Cowles [2], who wanted to measure the accuracy of the stock market predictors of his time. To do so, he generated several time series of predictions by simply drawing from a stack of cards labeled “positive,” “negative,” etc. Using these, he could determine if analysts were truly making predictions better than chance. Similarly, we judge analysts in reference to “pseudo-analysts,” which we generate such that they cover the same stock at the same time. So, against the pseudo-analysts, the real analyst has access to the same opportunities. Based on stock returns during the period, we compute the analyst’s percentile score against the pseudo-analysts. These percentile scores are a much fairer means of comparison than simply comparing raw portfolio returns. Moreover, our result allows for a more interpretable statement like “the analyst beat 60 % of pseudo-analysts,” rather than a statement from the portfolio method like “the analyst generated a + 5 % return” where interpretation is difficult without a great deal of context.

[1] Another approach is to use surveys, such as Institutional Investor’s annual awards, where experts, such as brokerage clients, are asked to rank analysts. They have been called “beauty contests” [2] since they can lack objectivity. Surveys can also be expensive and require expert participation.

2 Background

Since 1993, the *Wall Street Journal* (WSJ) has published its “Best on the Street” ranking of analysts. The methodology states: “For a stock rated a buy, a positive total return yielded a positive score on that stock, but a negative return produced a negative score. Similarly, for a stock rated sell, a negative total return yielded a positive score while a positive return resulted in a negative score. Hold recommendations did not affect the score” [1]. Consider the 2007 Pharmaceuticals Sector [7]:

> BEING ON THE RIGHT side of huge swings in small companies helped propel Jonathan Aschoff into the No.1 spot among pharmaceutical analysts . . . Mr. Aschoff . . . upgraded shares of Adolor(c) Corp. to buy from hold in early February, the day its shares surged 41 % . . . Mr. Aschoff downgraded the stock to sell in early September, the day the shares plunged 45 % . . . . He benefited from the methodology of this survey, which calculates returns from the closing the day before the recommendation change, scoring for both the 58 % return while he rated the stock a buy and a nearly 73 % decline during his sell recommendation.

Two shortcomings are evident. First, the previous day’s close is considered the starting point of the rating. In two instances, the largest portion of return came from stock movement occurring before the rating and, thus, does not reflect predictive ability. Second, unless the other analysts had access to stocks with similar “huge swings,” they would be unable to have a high rank, even if they were highly accurate on their lower opportunity stocks. As discussed in Sect. 1, analysts typically have limited control of their coverage, so the rankings can involve luck as much as skill.

StarMine has recognized some of these issues. The 2010 US award methodology states “The portfolio return is opportunity adjusted to facilitate a fair comparison of analyst performance regardless of their coverage universe.” The adjustment method is not specified, but there is strong evidence [8, 9] that they normalize by the volatility of the covered stocks. This can help, but volatility really measures noise rather than opportunity. Consider Fig. 1. Stocks A and B have standard deviations 5.2 and 5.7 %, respectively. StarMine’s methodology suggests stock B has greater opportunity. Yet, stock B essentially has noise around an upward path, while stock A has a sequence of returns that we might reasonably expect a good analyst could label so that an investor would know when to be long or short.

StarMine’s methodology also states “Holds invest one unit in the benchmark (i.e., for an excess return of zero).” Consider Fig. 2. Suppose an analyst covers one of the stocks and issues a hold. Regardless of whether the stock is C or D, his/her portfolio return would be identical under the StarMine methodology. Yet, the analyst would clearly be less correct about C than D (assuming a benchmark return of 0 %).

3 Data and Method

We use the Center for Research in Security Prices (CRSP)’s daily total return for each stock which includes not only price changes but all payouts (e.g., cash dividends). Values after delisting are also used, which prevents upward biases.

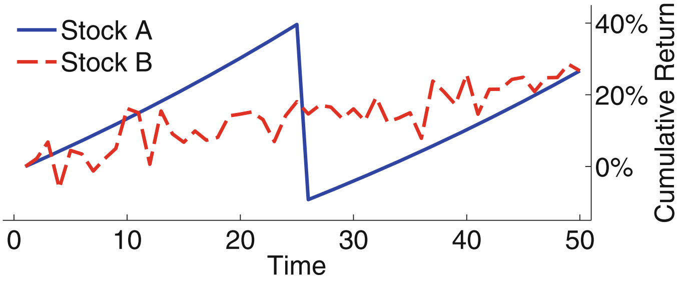


Fig. 1 Equal predictive opportunity

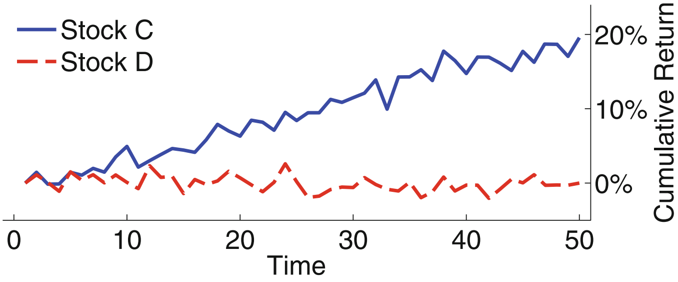


Fig. 2 Equal reward for Hold ratings

We use the Capital Asset Pricing Model (CAPM) to calculate abnormal return, which is the difference between a stock's actual and expected return. Abnormal return reflects the job of the analyst, which is usually not to predict if a stock will go up or down in absolute, but to predict its performance relative to the market or peers [5]. For a stock s over time period T, we calculate abnormal return $\hat{R}_{s,T}$ as

$$\hat{R}_{s,T} = R_{s,T} - R_{f,T} + \beta_{s,T} \cdot (R_{m,T} - R_{f,T}), \tag{1}$$

where $R_{s,T}$, $R_{m,T}$ and $R_{f,T}$ are the returns over time period T of stock s, the US market and a risk-free instrument, respectively. Market and risk-free returns are obtained from [4]. The sensitivity of stock s to the market is denoted by $\beta_{s,T}$, which is calculated immediately prior to time period T using the regression

$$r_{s,t} - r_{f,t} = \alpha + \beta_{s,T} \cdot (r_{m,t} - r_{f,t}) + \epsilon_t, \tag{2}$$

where $r_{s,t}$, $r_{m,t}$ and $r_{f,t}$ are 20-trading-day returns for the stock, market and a risk-free instrument at time t, respectively, and ϵ_t is the error at time t. The interval of 20 trading days is approximately 1 calender month and we regress over a 500-day period, so the regression is over 25 points.

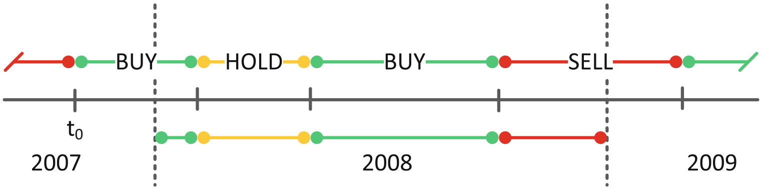


Fig. 3 An example timeline of analyst ratings for a single stock are shown on *top*. The lengths considered for the year's returns calculations are shown on *bottom*

For ratings, we use the I/B/E/S US dataset, which assigns each analyst a unique identifier that remains constant even if s/he switches firms. I/B/E/S also standardizes each firm's ratings into a five-level system of 1-Strong Buy, 2-Buy, 3-Hold, 4-Underperform, and 5-Sell. For simplicity, we use 1 and 2 as "Buy," 3 as "Hold," and 4 and 5 as "Sell." Once an analyst makes a rating, we consider it active until (1) the analyst issues a new rating for the stock, (2) a different analyst at the same firm issues a new rating (i.e., stock was reassigned), (3) a stop coverage is issued by the firm, (4) the stock is delisted, or (5) 250 trading days (approx. 1 calendar year) elapses.

WSJ and StarMine awards are annual. Correspondingly, we break our data into years. For a single stock s in a single year, suppose an analyst has ratings with time periods $T_1, T_2, ..., T_n$. We compute the cumulative abnormal return as

$$\bar{R}_s = \sum_{k=1}^{n} d_k \cdot \hat{R}_{s,T_k}, \tag{3}$$

where d_k corresponds to the rating at T_k, where Buy is $+1$, Hold is 0 and Sell is -1. As Fig. 3 exemplifies, returns are only counted within the measurement year.

Let $\mathbb{F}_y$ denote the set of all ratings of all analysts on all stocks where the ratings overlapped the measurement year, y. To evaluate a single analyst on a single stock over year y, we generate multiple pseudo-analysts where each analyst begins on the start date of the analyst's earliest rating that overlaps year y (t_0 in Fig. 3). We then randomly sample from $\mathbb{F}_y$ to generate a sequence of rating lengths until the end of the measurement year is reached or exceeded. Buy, Hold and Sell levels are subsequently applied by again sampling from $\mathbb{F}_y$. Thus, both the length and level distributions come directly from the real analyst population, matching our desire that the pseudo-analysts replicate the real analysts. This helps avoid biases (dis)favoring the real analysts, although it replicates their behavior, even if irrational (e.g., analysts tend to issue more buys than sells [6]).

For each pseudo-analyst, abnormal return is calculated in the same manner as the real analyst. We then compute a percentile value p_s, which is the fraction of pseudo-analysts that had lower abnormal return $\bar{R}_s$ than the real analyst. We compute a composite score over all stocks S in the analyst's coverage as

$$\bar{p}_s = \frac{\sum_{s \in S} \ell_s \cdot p_s}{\sum_{s \in S} \ell_s}, \tag{4}$$

where ℓ_s is the number of days stock s was covered during the year. Analysts can be ranked and compared by $\bar{p}_s$. As stated in Sect. 1, $\bar{p}_s$ is easily interpretable since it indicates how many random analysts the real analyst outperformed.

We recognize contention may exist over some aspects of our approach (e.g., returns should use a sector benchmark rather than a market benchmark, Strong Buy should be differentiated from Buy, etc.). These aspects can be altered for particular situations and tastes, yet, the shift to a MC approach is a significant structural improvement. For example, in the case of holds, lost opportunity is truly captured in the MC method because pseudo-analysts will have higher returns if a buy or sell rating was more appropriate. In the portfolio method, it is unclear if other analysts will have higher portfolio return since few others may be covering the same stock.

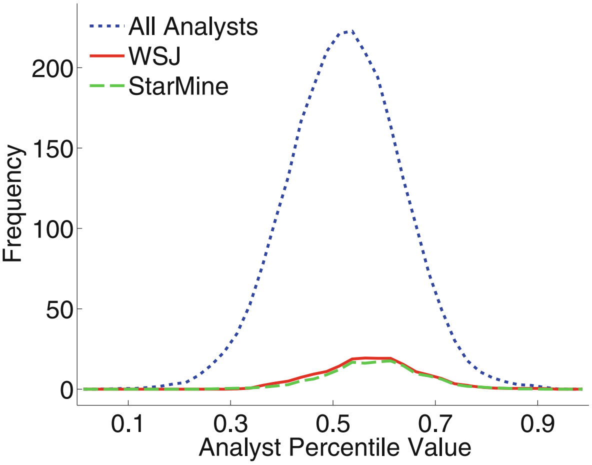
4 Experimental Results

WSJ and StarMine award analysts were collected from their respective websites for years 2001–2009 (award years 2002–2010). Analysts were manually linked to I/B/E/S using their names, firm at award time and textual descriptions. For each stock covered by each analyst in a given year, we generate 10,000 pseudo-analysts. Average percentile values for all real analysts and for WSJ and Starmine analysts are shown in Table 4. As can be seen in the All analysts column, the median percentile value is near 0.5, indicating that the average analyst tends to do no better than an average pseudo-analyst. With statistical significance at the 5 % level for all years using a Mann–Whitney U test, WSJ award analysts tend to do better than analysts without a WSJ award. The same is true of StarMine analysts. This is expected since an analyst with higher score under the MC approach would tend to have higher value

Table 1 Median Monte Carlo percentiles

Year	All analysts	WSJ	StarMine
2001	0.474	0.498	0.497
2002	0.487	0.545	0.543
2003	0.495	0.539	0.560
2004	0.505	0.560	0.556
2005	0.509	0.570	0.570
2006	0.500	0.540	0.551
2007	0.521	0.569	0.584
2008	0.494	0.523	0.528
2009	0.508	0.526	0.538

Fig. 4 Percentile distributions (2007 shown)



under the portfolio method. Yet, as seen in Fig. 4, the WSJ and StarMine awards do not capture many of the best analysts identified by the MC method. In analyzing the results, we find this may occur in some instances because an analyst with high MC score did not meet certain WSJ or StarMine requirements, such as covering a sufficient number of stocks in a particular industry. However, we find that it occurs frequently when an analyst has less opportunity to capture large returns. It also occurs when an analyst is outranked by other analysts with erroneous hold ratings but were not penalized for those rating under the WSJ and StarMine methodologies.

These results support our claim that the popular portfolio method does not properly capture analyst performance. We suggest the presented MC method alleviates the identified issues and offers a fairer representation of analyst accuracy.

References

1. Best on the Street (a special report): 2007 analysts survey—how the survey was conducted. Wall Str. J. R4 (2007). May 21
2. Cowles III, A., Can stock market forecasters forecast? Econometrica **1**(3), 309–324 (1933)
3. Emery, D.R., Li, X.: Are the wall street analyst rankings popularity contests? J. Financ. Quant. Anal. **44,** 411–437 (2009)
4. French, K.: Kenneth R. French-Data Library. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (2012). August 2013
5. Kadan, O., Madureira, L., Wang, R., Zach, T.: Industry recommendations: Characteristics, investment value, and relation to firm recommendations (2009)
6. McNichols, M., O'Brien, P.C.: Self-selection and analyst coverage. J. Account. Res. **35,** 167–199 (1997)
7. Rubenstein, S.: Best on the street (a special report): 2007 analysts survey—pharmaceuticals. Wall Str. J. R7 (2007). May 21
8. StarMine: 2003 North America industry analyst awards. http://excellence.thomsonreuters.com/award/starmine (2003). August 2013
9. StarMine: Coverage-relative rating. http://professionalhelp.starmine.com/help/analysts/analysts.phtml?page_set=coverage_relative_rating (2013). August 2013

Wake Topology for Steady Flow Past an Inclined Elliptic Cylinder

Peter J. S. Young

Abstract The steady flow of an incompressible viscous fluid past an elliptic cylinder with minor-to-major axis ratio of 0.2 and at incidence to the free stream is considered. Numerical results for Reynolds number up to 450 and inclination angle varying from 0° to 20° are presented which permit completion of a bifurcation diagram describing the wake topology behind the cylinder in terms of three regions: Region I with no separation; Region II with a single recirculatory region attached to the cylinder; and Region III with two recirculatory regions, one attached and one unattached.

1 Introduction

Numerical studies of the steady flow of an incompressible viscous fluid past an elliptic cylinder at inclination to the free stream have identified a wake topology defined by the presence of recirculatory regions behind the cylinder. Dennis and Young [3], in considering an elliptic cylinder with minor-to-major axis ratio of 0.2, identified three regions: Region I where there is no flow separation; Region II where there is a single recirculatory region attached to the cylinder; and Region III where there are two recirculatory regions, one attached and one unattached. Their results were summarised in a bifurcation diagram showing region boundaries as functions of Re and α for Re up to 40 and α varying from 0° to 90°. Sen et al. [6] confirmed this behaviour and predicted non-monotonic behaviour for the Region I–II boundary curve for Re in the range 184–205 and $0^\circ \leq \alpha \leq 5^0$.

The objective of the work reported here has been to extend numerical solutions of the flow past an inclined elliptic cylinder to higher Re to thereby complete the bifurcation diagram presented by [3]. This has been achieved through numerical solutions obtained for Re up to 450 and α varying from 0° to 20°.

This work does not reflect official work/views of the NCI Agency.

P. J. S. Young (✉)
NCI Agency, 2597 AK, The Hague, The Netherlands
e-mail: peter.young@ncia.nato.int

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_76

2 Problem Formulation

An elliptic cylinder with major and minor axis lengths of $2a$ and $2b$ is placed with its centre at the origin of Cartesian coordinates (x, y) and its major axis in the direction of x. The undisturbed stream has velocity U at angle α to the positive direction of x. The problem is formulated in the following elliptic coordinates (ξ, η):

$$x = \cosh\xi \cos(\eta + \alpha), \quad y = \sinh\xi \sin(\eta + \alpha) \tag{1}$$

The cylinder surface is associated with $\xi = \xi_0$ where $\xi_0 = \tanh^{-1}(b/a)$.

The flow field is described by the two-dimensional Navier Stokes equations. Using a stream function Ψ, vorticity ζ formulation, the non-dimensional Navier Stokes equations in elliptic coordinates are given by:

$$\frac{\partial^2\Psi}{\partial\xi^2} + \frac{\partial^2\Psi}{\partial\eta^2} + \tfrac{1}{2}(\cosh 2\xi - \cos 2(\eta + \alpha))\zeta = 0 \tag{2}$$

$$\frac{\partial^2\zeta}{\partial\xi^2} + \frac{\partial^2\zeta}{\partial\eta^2} = \frac{Re}{2}\left(\frac{\partial\Psi}{\partial\eta}\frac{\partial\zeta}{\partial\xi} - \frac{\partial\Psi}{\partial\xi}\frac{\partial\zeta}{\partial\eta}\right) \tag{3}$$

where $Re = (2Ua\cosh\xi_0)/\nu$ is the Reynolds number based on the half-length a of the major axis for the elliptic cylinder and ν is the dynamic viscosity of the fluid.

The boundary conditions, reflecting no slip on the cylinder surface, velocities approaching the free stream at far distances from the cylinder, and continuity with respect to the angular coordinate, are the following:

$$\Psi = \frac{\partial\Psi}{\partial\xi} = 0 \quad \text{when } \xi = \xi_0, \tag{4}$$

$$e^{-\xi}\frac{\partial\Psi}{\partial\xi} \to \tfrac{1}{2}\sin\eta, \quad \exp -\xi\frac{\partial\Psi}{\partial\eta} \to \tfrac{1}{2}\cos\eta, \quad \zeta \to 0 \quad \text{as } \xi \to \infty, \tag{5}$$

$$\Psi(\xi, \eta) = \Psi(\xi, \eta + 2\pi), \quad \zeta(\xi, \eta) = \zeta(\xi, \eta + 2\pi). \tag{6}$$

One difficulty in obtaining numerical solutions to this problem is the handling of the far wake with imposition of suitable boundary conditions for Ψ and ζ as $\xi \to \infty$. Imai [5] has obtained the leading terms of the asymptotic solution for this region, a key characteristic of which is ζ becomes singular in the far wake as $\xi \to \infty$. Outside of the far wake ζ is zero and the solution is governed by Laplace's equation for Ψ. The asymptotic problem has also been formulated by Dennis [1] using a Hermite polynomial expansion for ζ to eliminate the angular coordinate and reduce the problem to a set of ordinary differential equations. Young [7] obtained the leading term solutions to these equations and confirmed their agreement with [5].

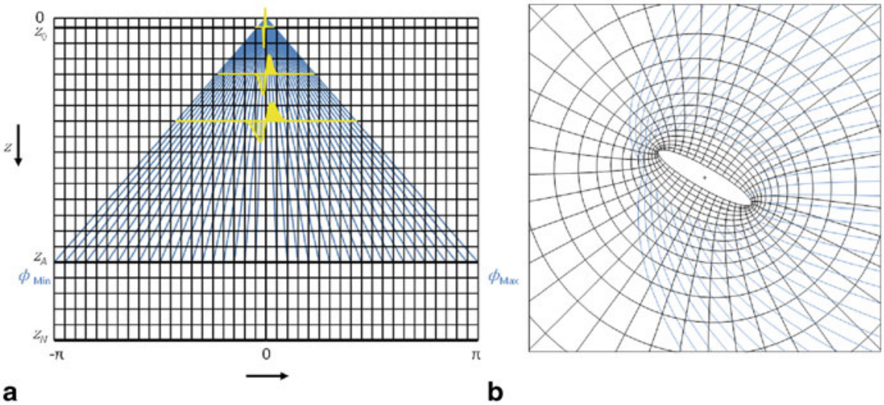


Fig. 1 Numerical grids: **a** (z, η) grid in *black* and (z, ϕ) grid in *blue* showing asymptotic nature of ζ as $z \to 0$; **b** (z, η) and (z, ϕ) grid lines in elliptic coordinates

The following variables are used to transform the Navier Stokes equations to enable the asymptotic solution for Ψ and ζ to be used:

$$z = \mathrm{e}^{-\xi/2}, \quad \phi = \frac{\eta}{2k}\mathrm{e}^{\xi/2} \quad \text{where } k = \sqrt{\frac{2}{R}}\,, \tag{7}$$

where R is the Reynolds number related to Re by $R = Re/\cosh \xi_0$. The variable ϕ takes on the range $(-\infty, \infty)$ as $z \to 0$, this corresponding to $\xi \to \infty$. The far wake, where the vorticity is nonzero, therefore becomes singular about $\eta = 0$.

3 Numerical Considerations

The Navier Stokes equations are approximated using central differences and solved using an iterative Gauss–Seidal approach. Two numerical grids are employed, Grid 1 in (z, η) space that covers the full flow domain and Grid 2 in (z, ϕ) space that maps onto the wake region behind the cylinder. These grids are shown in Fig. 1. Grid 2 has been adopted to capture the singular behaviour of the vorticity in the far wake as $z \to 0$, this being illustrated in Fig. 1a. Suitably transformed versions of the Navier Stokes equations are used for each of these grids.

Grid 1 is defined for the region $z_0 \le z \le z_N$ and $-\pi \le \eta \le \pi$ but with the Grid 2 subregion excluded. For Grid 1, the far field boundary condition for Ψ is applied at $z = z_0$ and the cylinder surface is specified at $z = z_N$. Grid 2 is defined for the region $z_0 \le z \le z_A$ and $\phi_{Min} \le \phi \le \phi_{Max}$, where the far field boundary conditions for ζ and Ψ are applied at $z = z_0$. The Grid 1–2 boundary closer to the cylinder is defined at $z = z_A$, and $\phi_{Max} = \pi/2kz_A$ (corresponding to $\eta = \pi$), with $\phi_{Min} = -\phi_{Max}$. The boundaries between the Grid 1 variable η and Grid 2 variable ϕ vary as a function of z given by $\eta_{\phi Max} = 2kz\phi_{Max}$, with $\eta_{\phi Min} = -\eta_{\phi Max}$. A numerical matching between Grids 1 and 2 is performed at these boundaries.

Boundary conditions are also required for Ψ and ζ at the cylinder surface $z = z_N$. Equation (4) is used for Ψ while a condition for ζ is obtained through a finite difference approximation at the cylinder surface based on Eqs. (2) and (4).

4 Results

Numerical results have been obtained for the elliptic cylinder with minor-to-major axis ratio of 0.2 for Re up to 450 and inclinations varying from 0° to 20°, with higher inclinations for Re up to 150. Validation of the numerical method has been performed through consideration of the symmetrical flow past a circular cylinder for Re up to 300, and for the asymmetric flow past an inclined elliptic cylinder for Re 1–40 and inclinations varying from 0° to 90°. The numbers of grid points taken for the z and η coordinates were $N = 160$ and $M = 160$. Results were also obtained using coarser grids in order to investigate grid dependence of the results.

Numerical results obtained for the symmetric flow past a circular cylinder favourably compare with Fornberg [4], e.g. present results using a 160×160 grid (for $0 \le \eta \le \pi$) at $Re = 300$ give $C_D = 0.723$ compared with $C_D = 0.729$ from [4], and a wake bubble length of 40.61 compared with 40.4 estimated from Fig. 10 of [4].

A comparison of results for the flow past an elliptic cylinder at $Re = 20$ and 40 and α varying from 0° to 90° was made with [3, 6]. At $Re = 20$ there is good agreement across all results with discrepancies in the order of 1%. Greater discrepancies were found for results at $Re = 40$. Drag coefficient results were in good agreement with [6], discrepancies being 1–2 %. A comparison of lift coefficient results with [6] found discrepancies of 8 % at $\alpha = 10°$, this reducing to 1 % for $\alpha \ge 40°$. A comparison of drag coefficient results with [3] found discrepancies in the order of 1 % for $\alpha \le 30°$, this increasing to 14 % at $\alpha = 90°$. For the lift coefficient, differences with [3] varied between 3 % at $\alpha = 30°$ to 22 % at $\alpha = 80°$. The streamline patterns obtained by [6] for $Re = 40$ and $\alpha \ge 45°$, when 2 recirculatory regions are present, always had the upper bubble attached to the cylinder. In contrast, [3] found for $\alpha \ge 57°$ the streamline pattern transitioned to having the lower bubble attached. The results obtained in the present study are consistent with [6] on this.

Completion of the bifurcation diagram required determination of Re at which separation first occurs for the elliptic cylinder at 0° incidence. Solutions were obtained at $Re = 190$ (no separation) and 195 (separation) from which initial separation is estimated at $Re = 192$. This is in close agreement with Dennis and Chang [2] who estimated initial separation to occur at some Re near but less than 200, and [6] with prediction $Re = 184.75$. Further numerical solutions were obtained at various Re and α to permit completion of the bifurcation diagram, which is presented in Fig. 2. The immediate observation from this is the non-monotonic behaviour for $Re \ge 192$ and $0° \le \alpha \le 17°$. Such behaviour was predicted by [6] for the Region I–II curve shown in Fig. 2, this being based on a linearity property of eddy length with Re.

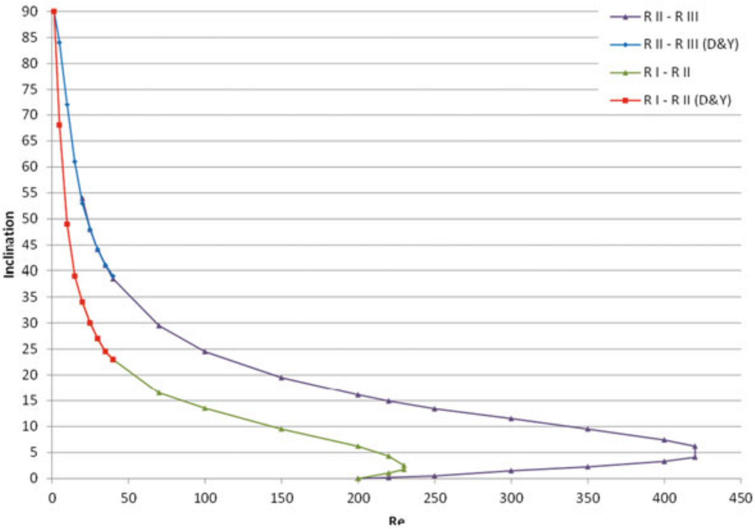


Fig. 2 Bifurcation diagram showing separation of Regions in (Re, α) space

Given the symmetric flow at some Re just above 192, introduction of a small angle of incidence results in the flow transitioning to Region III topology with a single attached upper separation bubble complemented by a second unattached lower bubble. The sizes of these recirculatory regions are found to decrease with increasing α until the unattached bubble disappears thereby giving a Region II topology. The attached bubble continues to decrease in size with further increases in α until it also disappears yielding Region I topology. Further increases in α then lead to the behaviour reported by [3]. This behaviour is illustrated in Fig. (3) for $Re = 220$. The range of α over which this occurs decreases with increasing Re until, at some Re in the range 230–240, the wake topology transitions to the flow being separated for all α. The Region II–III curve in Fig. 2 closes at some Re just greater than 420, above which the flow pattern consists of 2 recirculatory regions for all $\alpha > 0°$.

Results for the lift and drag coefficients obtained for Re in the range 200–450 and $0° \leq \alpha \leq 20°$ are shown in Figs. 4 and 5.

5 Conclusions

Numerical results for the steady flow past an elliptic cylinder have been obtained for Re up to 450 at various inclinations. These results have permitted completion of the bifurcation diagram presented by [3], which characterises flow solutions on the basis of recirculatory regions behind the cylinder. The trend of the findings are in agreement with predictions by [6] of a non-monotonicity in the relationship between incidence angle and Re for initial separation. The non-monotonic behaviour for

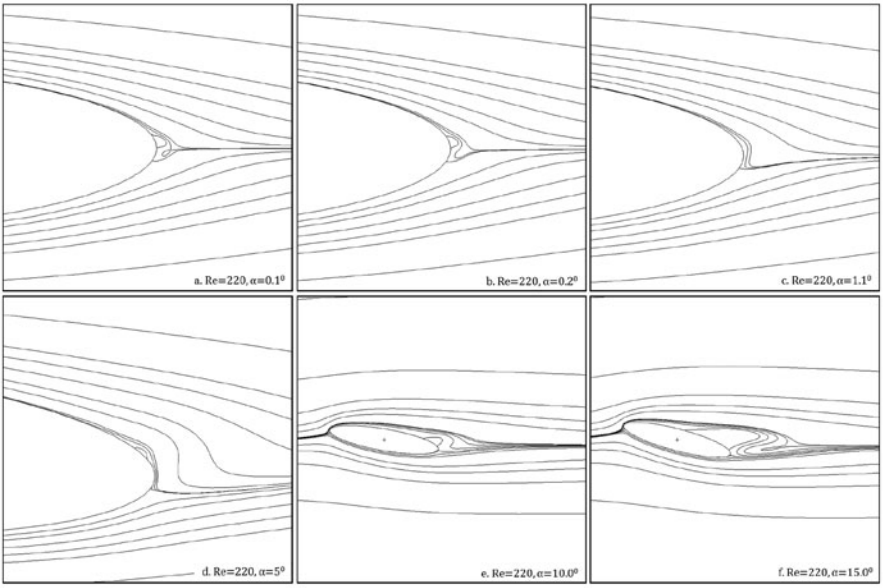


Fig. 3 Streamlines for $Re = 220$, inclinations: **a** $\alpha = 0.1°$ **b** $\alpha = 0.2°$ **c** $\alpha = 1.1°$, **d** $\alpha = 5.0°$, e. $\alpha = 10.0°$, f. $\alpha = 15.0°$

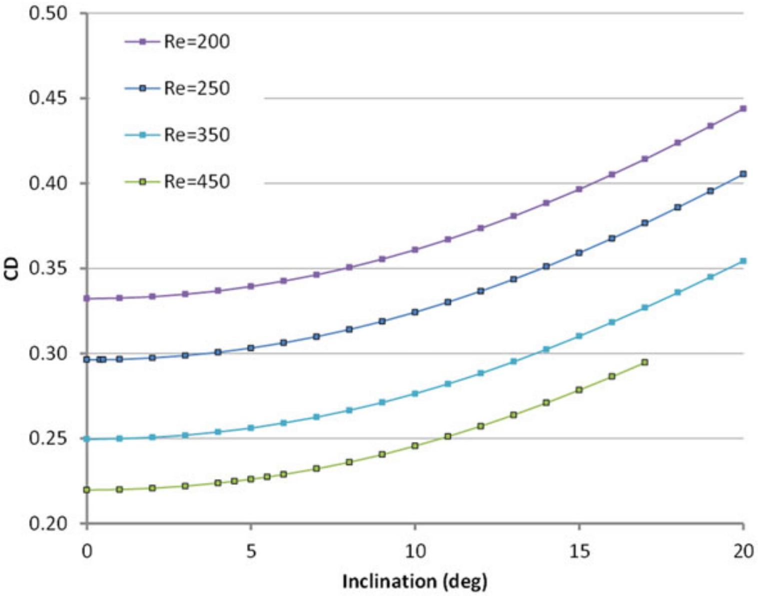


Fig. 4 Drag coefficients

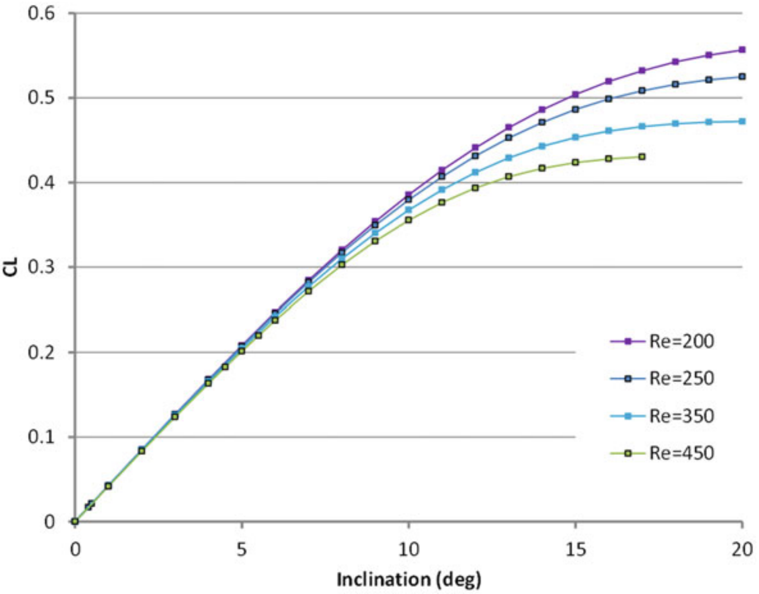


Fig. 5 Lift coefficients

$Re \geq 192$ with incidence angle increasing from 0° is of interest, particularly in how the separated bubble initially decreases in size and disappears before reappearing again.

References

1. Dennis, S.C.R.: A numerical method for calculating two-dimensional wakes, AGARD Conference Proceedings #60 on Numerical Methods for Viscous Flows (1967)
2. Dennis, S.C.R., Chang, G.Z.: Numerical integration of the Navier Stokes equations in two dimensions. Mathematics Research Center, University of Wisconsin, Technical Summary Report 859, p. 89 (1969)
3. Dennis, S.C.R., Young, P.J.S.: Steady flow past an elliptic cylinder inclined to the stream. J. Eng. Math. **47,** 101–120 (2003)
4. Fornberg, B.: Steady viscous flow past a circular cylinder up to Reynolds Number 600. J. Comput. Phys. **61**(2), 297–320 (1985)
5. Imai, I.: On the asymptotic behaviour of viscous fluid flow at a great distance from a cylindrical body, with special reference to Filon?s paradox. Proc. R. Soc. Lond. A **208,** 487–516 (1951)
6. Sen, S., Mittal, S., Biswas, G.: Steady separated flow past elliptic cylinders using a stabilized finite-element method. CMES **86**(1), 1–26 (2012)
7. Young, P.J.S.: Steady asymmetric flow of a viscous fluid past a cylinder, Ph.D. Thesis, University of Western Ontario, London, Ontario, Canada, p. 139 (1989)

Leading Unstable Linear Systems to Chaos by Chaos Entanglement

Hongtao Zhang, Xinzhi Liu and Xianguo Li

Abstract Chaos entanglement is a new approach to systematically generate chaotic dynamics by entangling two or multiple stable linear systems with periodic nonlinear coupling functions such that each of them evolves in a chaotic manner. In this study, chaos entanglement is extended to unstable linear systems by introducing a well-defined bound function to guarantee the boundedness of each unstable linear system. A novel 6-scroll attractor is obtained by entangling three identical unstable linear systems with sine function. It is verified that this attractor possesses a positive Lyapunov exponent and its trajectories are bounded. The Lyapunov spectra and bifurcation diagram reveal the chaotic behaviors of this new attractor.

1 Introduction

Chaos phenomena, characterized by the so-called "butterfly effect", have been found in many fields such as physics, biology, philosophy, economics, and engineering. Significant attention is attracted due to the deterministic characteristic and unpredictable essence of chaotic systems. Specifically, since the pioneering work by Pecora and Carroll [16], chaotic systems have become useful tools in engineering applications, for instance, the carrier for secure communication, the random bit generator, and the chaotic radar (see [5, 9, 15, 18, 21, 28] and references therein).

Methods to construct new chaotic attractors mainly focus on extending and generalizing existing chaotic systems such as Lorenz system [11], Chua's circuit [14], and Mackey-Glass equation [13]. Starting from Lorenz system, a simplified mathematical model for atmospheric convection, Chen attractor was presented [3]. A

H. Zhang (✉) · X. Liu · X. Li
University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
e-mail: h15zhang@gmail.com

X. Liu
e-mail: xzliu@uwaterloo.ca

X. Li
e-mail: xianguo.li@uwaterloo.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_77

general canonical form of Lorenz system was established [2]. Multi-scroll Chen attractor was generated [10]. Starting from Chua's circuit, the first real-world chaotic circuit, multi-scroll chaotic/hyperchaotic (with more than one positive Lyapunov exponent) attractors were reported such as a n-double scroll attractor [19], a multi-grid attractor [24], a multi-scroll and hypercube attractor [22], a 3-D multi-scroll chaotic attractor [12], and a multi-folded torus chaotic attractor [25]. Starting from Mackey–Glass equation, a physiological control system with time delay, a modified attractor with a piecewise nonlinearity is obtained [20]. More results were presented such as a n-scroll chaotic attractor from a delay differential equation [23], the simplest delay differential equation to generate chaotic attractors [17], a family of novel chaotic/hyperchaotic attractors from a first-order delay differential equation with sine function [26], and a delay dynamical system with hyperbolic tangent function [1]. Also, some new technologies were adapted to construct new chaotic attractors, for instance, nonautonomous techniques [7], switching approaches [8], and fractional differential equations [6]. A considerable success has been achieved in creating new chaotic attractors by above methods. However, some problems arise due to the similarity of relative attractors, for instance, the security of chaos-based secure communication. Chaos entanglement was presented by [27] as a bridge to connect linear systems to chaos, by entangling two or multiple stable linear systems to form an artificial chaotic system/network such that each of them evolves in a chaotic manner. Variable strange attractors and abundant dynamical behaviors reveal that this approach possesses the potential to create various desired chaotic attractors without similarity. Unfortunately, it is validated only applicable to stable linear systems. For unstable linear systems, it fails as the periodic entanglement functions could not guarantee the boundedness of the entire system, which is a necessary condition for generating chaos.

The objective of this study is to further explore the potential of chaos entanglement, extending it to unstable linear systems. First, a piecewise bound function is well defined and introduced to each unstable linear system. Furthermore, a 6-scroll strange attractor is achieved by entangling three identical unstable linear systems with sine function. The remainder of this paper is organized as follows. In Sect. 2, the new chaotic attractor is described in detail. Lyapunov spectra and bifurcation are analyzed in Sect. 3. Conclusions are given in Sect. 4.

2 New Attractor

Chaos entanglement is to entangle two or multiple linear systems to form an artificial chaotic system/network such that each subsystem evolves in a chaotic manner. There are two conditions required: one is that each subsystem should be stable while the other is that the entanglement function is periodic [27]. Consider three identical unstable linear subsystems as follows,

$$\dot{x}_i(t) = k_1 x_i(t), \quad for \quad i = 1, 2, 3, \tag{1}$$

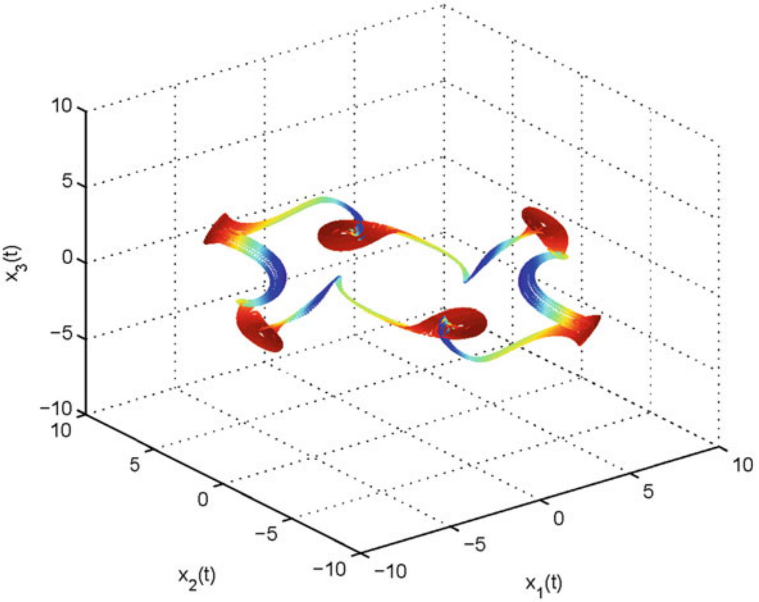


Fig. 1 Chaotic attractor given by system (3) with $k_1 = 0.12$, $k_2 = -3$, $x_0 = 5$, and $b_i = 5$ $(i = 1, 2, 3)$ starting from $(1, -2, 3)$

where $x_i(t)$ $(i = 1, 2, 3)$ is the state variable. For $k_1 > 0$, all three subsystems are unstable. A bound function is defined as,

$$g(x) = \begin{cases} (k_2 - k_1)(x - x_0), & x > x_0; \\ 0, & |x| \le x_0; \\ (k_2 - k_1)(x + x_0), & x < -x_0. \end{cases} \tag{2}$$

where x_0 is a positive constant, k_2 is a negative real number. Adding this bound function to each subsystem and entangling them by sine function gives

$$\begin{cases} \dot{x}_1(t) = k_1 x_1(t) + g(x_1(t)) + b_1 \sin(x_2(t)) \\ \dot{x}_2(t) = k_1 x_2(t) + g(x_2(t)) + b_2 \sin(x_3(t)), \\ \dot{x}_3(t) = k_1 x_3(t) + g(x_3(t)) + b_3 \sin(x_1(t)) \end{cases} \tag{3}$$

where b_i $(i = 1, 2, 3)$ is the entanglement coefficient and sin (.) is the entanglement function. For $k_1 = 0.12$, $k_2 = -3$, $x_0 = 5$, and $b_i = 5$ $(i = 1, 2, 3)$, a 6-scroll attractor is obtained as shown in Fig. 1. Numerical computation confirms that this chaotic system possesses a positive Lyapunov exponent with $\lambda = 0.2743$.

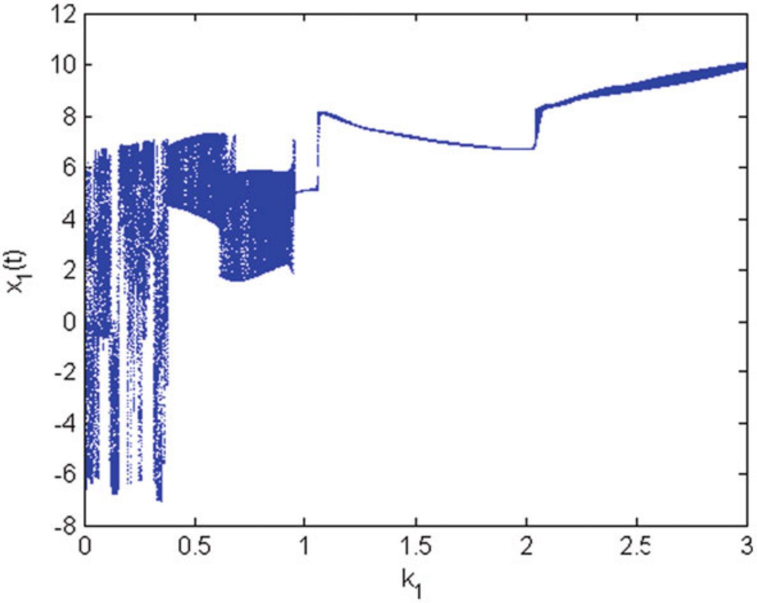


Fig. 2 The bifurcation diagram of x_1 versus k_1 with k_1 ranging in $[0,\ 3]$, $k_2 = -3$, $x_0 = 5$, and $b_i = 5\ (i = 1, 2, 3)$

3 Bifurcation and Lyapunov Spectra

The Lyapunov exponents of system (3) are calculated as $\lambda_1 = 0.2743$, $\lambda_2 = 0$, and $\lambda_3 = -4.4946$ for $k_1 = 0.12$, $k_2 = -3$, $x_0 = 5$, and $b_i = 5\ (i = 1, 2, 3)$. Since the largest one is positive, combining with its boundedness, it could be said that system (3) is chaotic. Furthermore, its Lyapunov dimension (Kaplan–Yorke dimension) is calculated as

$$D_{KY} = k + \sum_{i=1}^{j} \frac{\lambda_i}{\lambda_{k+1}} = 2 + \frac{\lambda_1 + \lambda_2}{|\lambda_3|} = 2.0610,$$

where k is the maximum integer such that the sum of the k largest exponents is still nonnegative. If $k = 0$, let $D_{KY} = 0$.

Next, we study the bifurcation diagram and Lyapunov spectra of system (3). Fix $k_2 = -3$, $x_0 = 5$, and $b_i = 5\ (i = 1, 2, 3)$ and let k_1 vary in $[0,\ 3]$. Figure 2 is the bifurcation diagram of x_1 versus k_1. The Lyapunov spectra is shown in Fig. 3. The uppermost curve (the blue) denotes its maximal Lyapunov exponent, which is an indication of chaos. When this blue curve goes into the upper half plane, system (3) turns into chaos. The chaotic regions mainly distribute in $(0,\ 0.3)$. In addition, it can be observed whenever the uppermost curve becomes positive, the middle one (the green) rapidly approaches to zero, which is in coincidence with the fact that at least one Lyapunov exponent vanishes (equal to 0) if the trajectory of an attractor does not contain a fixed point [4].

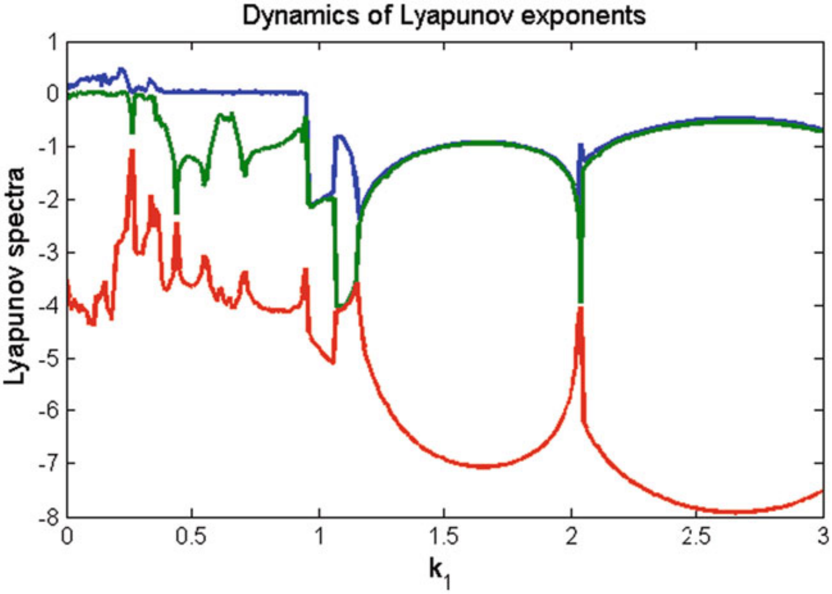


Fig. 3 The Lyapunov spectra with k_1 ranging in [0, 3], $k_2 = -3$, $x_0 = 5$, and $b_i = 5$ $(i = 1, 2, 3)$

4 Conclusion

In this present study, chaos entanglement has been extended to unstable linear systems by introducing a piecewise bound function to guarantee the boundedness of each system. A 6-scroll strange attractor has been achieved. Numerical computation has confirmed that this attractor possesses a positive Lyapunov exponent. Furthermore, its chaotic behaviors have been observed by the bifurcation diagram and Lyapunov spectra. To further improve this method, there are two outstanding issues open to study now. One is to find out the necessary and sufficient conditions for chaos entanglement while the other is to generate hyper-chaos by chaos entanglement. Specifically, to further explore possible connections between linear systems and chaotic attractors will be our future work.

Acknowledgments The financial support for this study is provided by the Ontario Research Fund-Research Excellence (ORF-RE) program via contract number # RE-02-019 and the Natural Sciences and Engineering Research Council of Canada (NSERC) via the Discovery Grant.

References

1. Banerjee, T., Biswas, D.: Theory and experiment of a first-order chaotic delay dynamical system. Int. J. Bifurcation Chaos **23**(06), 1330020 (2013)
2. Čelikovskỳ, S., Chen, G.: On a generalized Lorenz canonical form of chaotic systems. Int. J. Bifurcation Chaos **12**(4), 1789–1812 (2002)
3. Chen, G., Ueta, T.: Yet another chaotic attractor. Int. J. Bifurcation Chaos **9**, 1465—1466 (1999)
4. Haken, H.: At least one Lyapunov exponent vanishes if the trajectory of an attractor does not contain a fixed point. Phys. Lett. A **94**(2), 71–72 (1983)
5. Kanter, I., Aviad, Y., Reidler, I., Cohen, E., Rosenbluh, M.: An optical ultrafast random bit generator. Nat. Photonics **4**(1), 58–61 (2010)
6. Li, C., Chen, G.: Chaos in the fractional order Chen system and its control. Chaos Solitons Fractals **22**(3), 549–554 (2004)
7. Li, Y., Liu, X., Zhang, H.: Dynamical analysis and impulsive control of a new hyperchaotic system*. Math. Comput. Model. **42**(11–12), 1359–1374 (2005)
8. Liu, X., Teo, K., Zhang, H., Chen, G.: Switching control of linear systems for generating chaos. Chaos Solitons Fractals **30**(3), 725–733 (2006)
9. Liu, X., Shen, X., Zhang, H.: Intermittent impulsive synchronization of chaotic delayed neural networks. Differ. Equ. Dyn. Syst. **19**(1–2), 149–169 (2011)
10. Liu, X., Shen, X., Zhang, H.: Multi-scroll chaotic and hyperchaotic attractors generated from Chen system. Int. J. Bifurcation Chaos **22**(02), 1250033 (2012)
11. Lorenz, E.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**(2), 130–141 (1963)
12. Lü, J., Han, F., Yu, X., Chen, G.: Generating 3-d multi-scroll chaotic attractors: a hysteresis series switching method. Automatica **40**(10), 1677–1687 (2004)
13. Mackey, M., Glass, L.: Oscillation and chaos in physiological control systems. Science **197**(4300), 287–289 (1977)
14. Matsumoto, T.: A chaotic attractor from Chua's circuit. IEEE Trans. Circuits Syst. **31**(12), 1055–1058 (1984)
15. Nguimdo, R.M., Colet, P., Larger, L., Pesquera, L.: Digital key for chaos communication performing time delay concealment. Phys. Rev. Lett. **107**(3), 034103 (2011)
16. Pecora, L., Carroll, T.: Synchronization in chaotic systems. Phys. Rev. Lett. **64**(8), 821–824 (1990)
17. Sprott, J.: A simple chaotic delay differential equation. Phys. Lett. A **366**(4–5), 397–402 (2007)
18. Sunada, S., Harayama, T., Davis, P., Tsuzuki, K., Arai, K., Yoshimura, K., Uchida, A.: Noise amplification by chaotic dynamics in a delayed feedback laser system and its application to nondeterministic random bit generation. Chaos: Interdiscip. J. Nonlinear Sci. **22**, 4754872 (2012)
19. Suykens, J., Vandewalle, J.: Generation of n-double scrolls (n = 1, 2, 3, 4, . . .). IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **40**(11), 861–867 (1993)
20. Tamasevicius, A., Mykolaitis, G., Bumeliene, S.: Delayed feedback chaotic oscillator with improved spectral characteristics. Electron. Lett. **42**(13), 736–737 (2006)
21. Willsey, M.S., Cuomo, K.M., Oppenheim, A.V.: Selecting the Lorenz parameters for wideband radar waveform generation. Int. J. Bifurcation Chaos **21**(09), 2539–2545 (2011)
22. Yalçin, M.: Multi-scroll and hypercube attractors from a general Jerk circuit using Josephson junctions. Chaos Solitons Fractals **34**(5), 1659–1666 (2007)
23. Yalçin, M., Özoguz, S.: n-scroll chaotic attractors from a first-order time-delay differential equation. Chaos: Interdiscip. J. Nonlinear Sci. **17**(3), 033112 (2007)
24. Yalçin, M., Suykens, J., Vandewalle, J., Ozoguz, S.: Families of scroll grid attractors. Int. J. Bifurcation Chaos **12**(1), 23–42 (2002)
25. Yu, S., Lu, J., Chen, G.: Multifolded torus chaotic attractors: design and implementation. Chaos: Interdiscip. J. Nonlinear Sci. **17**(1), 013118 (2007)

26. Zhang, H., Liu, X., Shen, X., Liu, J.: A family of novel chaotic and hyperchaotic attractors from delay differential equation. Dyn. Continuous Discrete Impulsive Syst. Series B: Appl. Algorithms **19**(03), 411–430 (2012)
27. Zhang, H., Liu, X., Shen, X., Liu, J.: Chaos entanglement: a new approach to generate chaos. Int. J. Bifurcation Chaos **23**(05), 1330014 (2013)
28. Zhang, H., Liu, X., Shen, X., Liu, J.: Intermittent impulsive synchronization of hyperchaos with application to secure communication. Asian J. Control **15**(6), 1686–1699 (2013)

Impulsive Control and Synchronization of Spatiotemporal Chaos in the Gray–Scott Model

Kexue Zhang, Xinzhi Liu and Wei-Chau Xie

Abstract This chapter investigates the impulsive control and synchronization problem of spatiotemporal chaos in the Gray–Scott model. Based on the Lyapunov function method, a class of pinning impulsive controller is designed to stabilize and synchronize the spatiotemporal chaos in the Gray–Scott model. The approach allows us to analyze the stability and synchronization problem of other spatiotemporal chaotic systems with the same structure. Numerical simulations are provided to illustrate the theoretical results.

1 Introduction

The theory of impulsive differential equations (IDEs) has been a very active research area for the past decades, since IDEs provide a framework for us to handle the mathematical modeling of many real-world dynamical systems which subject to abrupt changes of the states at some discrete times. Based on the theory of IDEs, the impulsive control method, the idea of which is to control the states of a system by using small impulses at discrete moments, has been widely used in various control problems.

Since the impulsive control method has been successfully used for systems modeled by ordinary differential equations, it is natural for us to consider the impulsive control problems of systems described by partial differential equations. In [2], the

K. Zhang (✉) · X. Liu
Department of Applied Mathematics, University of Waterloo,
Waterloo, ON N2L 3G1, Canada
e-mail: k57zhang@uwaterloo.ca

X. Liu
e-mail: xzliu@uwaterloo.ca

W.-C. Xie
Department of Civil and Environmental Engineering, University of Waterloo,
Waterloo, ON N2L 3G1, Canada
e-mail: xie@uwaterloo.ca

M. G. Cojocaru et al. (eds.), *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science,* Springer Proceedings in Mathematics & Statistics 117,
DOI 10.1007/978-3-319-12307-3_78

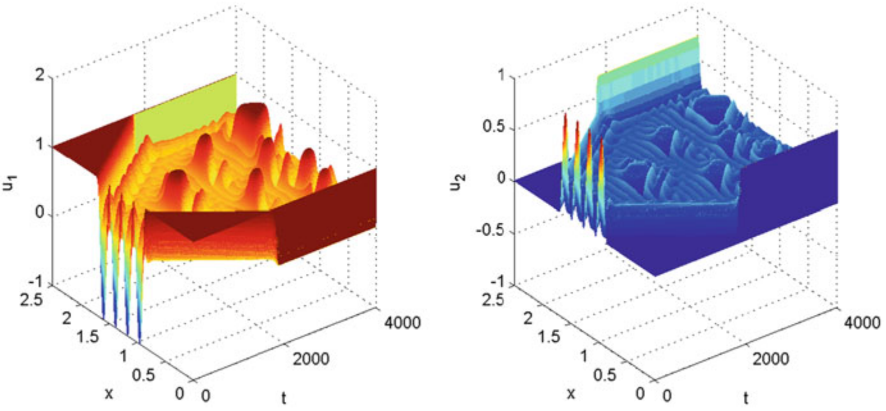


Fig. 1 Spatiotemporal evolutions of $u_1(t,x)$ and $u_2(t,x)$

impulsive control method was first introduced to the control problem of spatiotemporal chaotic systems represented by partial differential equations. Impulsive control problem of Kuramoto–Sivashinsky equation and impulsive synchronization problem of Gray–Scott model have been reported. Inspired by the pinning impulsive control method in [5], we shall extend the application of the pinning impulsive control method to the synchronization problem of the Gray–Scott model.

The Gray–Scott model is one of the typical reaction–diffusion systems which has a wide variety of spatiotemporal structures [1]:

$$\begin{aligned}\frac{\partial u_1}{\partial t} &= -u_1u_2^2 + a(1-u_1) + d_1\nabla^2 u_1,\\ \frac{\partial u_2}{\partial t} &= u_1u_2^2 - (a+b)u_2 + d_2\nabla^2 u_2,\end{aligned} \tag{1}$$

where u_1 and u_2 are the concentrations of chemical species U_1 and U_2, respectively, a is the flow rate, $a+b$ is the removal rate of U_2 from the reaction, and d_1 and d_2 are the diffusion coefficients of the two species.

In this chapter, we consider the one-dimensional version of the Gray–Scott model with $a = 0.028$, $b = 0.053$, $d_1 = 2 \times 10^{-5}$, and $d_2 = 10^{-5}$. Since $E_0 = (1,0)$ is a trivial steady state, it is necessary to add certain perturbation to it to obtain nontrivial pattern from the initial state $(1,0)$. The initial conditions are chosen to be $(u_1(0,x), u_2(0,x))^T = (1,0)^T$ with strong perturbations in the center region, and the periodic boundary conditions are given by $u_1(t,0) = u_1(t,L) = 1$ and $u_2(t,0) = u_2(t,L) = 0$. The spatiotemporal chaotic evolutions of the 1-D system (1) are shown in Fig. 1. For more details about the Gray–Scott model and its chaotic dynamics, please refer to [3] and [6].

2 Impulsive Synchronization of Spatiotemporal Chaos

In this section, we shall discuss the impulsive synchronization of one-dimensional Gray–Scott model with another identical system starting from different initial states.

Let the following one-dimensional Gray–Scott model serve as the drive system:

$$\begin{cases} \frac{\partial u_1}{\partial t} = -u_1u_2^2 + a(1-u_1) + d_1\frac{\partial^2 u_1}{\partial x^2}, \\ \frac{\partial u_2}{\partial t} = u_1u_2^2 - (a+b)u_2 + d_2\frac{\partial^2 u_2}{\partial x^2}, \\ \mathbf{u}(0,x) = \mathbf{u}_0(x), \;\; x \in [0,L], \\ \mathbf{u}(t,0) = \mathbf{u}(t,L) = \mathbf{h}(t), \;\; t \in \mathbb{R}^+, \end{cases} \tag{2}$$

where $\mathbf{u}(t,x) = (u_1(t,x), u_2(t,x))^T$, and the response system is given by

$$\begin{cases} \frac{\partial v_1}{\partial t} = -v_1v_2^2 + a(1-v_1) + d_1\frac{\partial^2 v_1}{\partial x^2}, \; t \neq t_k, \\ \frac{\partial v_2}{\partial t} = v_1v_2^2 - (a+b)v_2 + d_2\frac{\partial^2 v_2}{\partial x^2}, \; t \neq t_k, \\ \Delta\mathbf{v}(t,x) = I_k(\mathbf{e}(t,x)), \;\; t = t_k, \; x \in [0,L], \; k = 1,2,..., \\ \mathbf{v}(0,x) = \mathbf{v}_0(x), \;\; x \in [0,L], \\ \mathbf{v}(t,0) = \mathbf{v}(t,L) = \mathbf{h}(t), \;\; t \in \mathbb{R}^+, \end{cases} \tag{3}$$

where a, b, d_1, and d_2 are chosen as in the previous section, $L = 2.5$, $\mathbf{v}(t,x) = (v_1(t,x), v_2(t,x))^T$, $\mathbf{u}_0$ and $\mathbf{v}_0$ are different initial conditions, $\mathbf{h}(t)$ is the periodic boundary condition for both systems. Since the Gray–Scott model exhibits chaotic behaviors, the same Gray–Scott systems will evolve differently if they have different initial conditions.

In (3), $I_k : \mathbb{R}^2 \to \mathbb{R}^2$, $\Delta\mathbf{v}(t,x) = \mathbf{v}(t^+,x) - \mathbf{v}(t^-,x)$, where $\mathbf{v}(t^+,x)$ and $\mathbf{v}(t^-,x)$ denote the right limit and left limit of $\mathbf{v}(t,x)$ at t, respectively. $\mathbf{e}(t,x) = \mathbf{u}(t,x) - \mathbf{v}(t,x)$ denotes the error state of the drive system and response system. The sequence $\{t_k\}$ satisfies $0 = t_0 < t_1 < t_2 < ... < t_n < ...$, and $\lim_{n\to\infty} t_n = \infty$.

According to (2) and (3), the error system will be given

$$\begin{cases} \frac{\partial e_1}{\partial t} = -u_1u_2^2 + v_1v_2^2 - ae_1 + d_1\frac{\partial^2 e_1}{\partial x^2}, \; t \neq t_k, \\ \frac{\partial e_2}{\partial t} = u_1u_2^2 - v_1v_2^2 - (a+b)e_2 + d_2\frac{\partial^2 e_2}{\partial x^2}, \; t \neq t_k, \\ \Delta\mathbf{e}(t,x) = I_k(\mathbf{e}(t,x)), \;\; t = t_k, \; x \in [0,L], \; k = 1,2,..., \\ \mathbf{e}(0,x) = \mathbf{e}_0(x), \;\; x \in [0,L], \\ \mathbf{e}(t,0) = \mathbf{e}(t,L) = \mathbf{0}, \;\; t \in \mathbb{R}^+, \end{cases} \tag{4}$$

where $\mathbf{e}_0(x) = \mathbf{u}_0(x) - \mathbf{v}_0(x)$.

Definition 1 Suppose that $\mathbf{u}(t,x) : \mathbb{R}^+ \times [0,L] \to \mathbb{R}^m$ for some $m > 0$, where $\mathbf{u}$ is of class $\mathcal{L}_2[0,L]$ with respect to x. Then $\|\cdot\|_2$ is defined by $\|\mathbf{u}(t,x)\|_2 := \left[\int_0^L \|\mathbf{u}(t,x)\|^2 dx\right]^{1/2}$, where $\|\cdot\|$ is the Euclidean norm.

Definition 2 We say that synchronization of the drive system (2) and the response system (3) are achieved under impulsive controller $\{t_k, I_k\}$ if $\lim_{t\to\infty} \|\mathbf{u}(t,x) - \mathbf{v}(t,x)\|_2 = 0$.

Clearly, exploring the synchronization of the two systems (2) and (3) is equivalent to investigating the attractive property of the error states $\lim_{t\to\infty} \|\mathbf{e}(t,x)\|_2 = 0$.

In order to force the response system (3) to synchronize with the drive system (2), we design the following impulsive controller:

$$I_k(\mathbf{e}(t_k,x)) = \begin{cases} -qe_i(t_k,x), & i = \mathcal{D}_k, \\ 0, & i \neq \mathcal{D}_k, \end{cases} \tag{5}$$

where the constant $q \in (0,1]$ is the impulsive strength to be designed, and the index $\mathcal{D}_k$ is defined as follows: for the impulsive instant t_k, one can reorder the error states $e_1(t_k,x)$ and $e_2(t_k,x)$ such that $\|e_{j_1}(t_k,x)\|_2 \geq \|e_{j_2}(t_k,x)\|_2$, then the index $\mathcal{D}_k$ is defined as $\mathcal{D}_k = j_1$. We can see that the controller is only added to one state of the response system (3) at each impulsive instant t_k.

In [2], sufficient conditions about uniform impulsive controller are designed, which require an upper bound for each impulsive interval. In order to improve these sufficient conditions, we introduce the following definition.

Definition 3 ([4] **Average Impulsive Interval**) The average impulsive interval of impulsive sequence $\zeta = \{t_k\}$ is less than T_a, if there exist a positive integer N_0 and a positive number T_a, so that $N_\zeta(T,t) \geq \frac{T-t}{T_a} - N_0, \quad \forall T \geq t \geq 0$, where $N_\zeta(T,t)$ denotes the number of impulsive times of the impulsive sequence ζ in the time interval (t,T).

According to this definition, there is no requirement on the upper bound of each impulsive interval, which is necessary for the impulsive control scheme in [2].

Now we are in the position to introduce the main result to guarantee the synchronization of the drive system (2) and the response system (3).

Theorem 1 *Suppose the average impulsive interval of the impulsive sequence $\zeta = \{t_k\}$ is less than T_a. Let $\rho = 1 - q(q-2)/2$, and $\beta = 4\beta_2\sqrt{\beta_1^2 + 4\beta_2^2} - 2a$, where $\beta_i := \max\{\sup_{t\in\mathbb{R}^+} |u_i(t,x)|, \sup_{t\in\mathbb{R}^+} |v_i(t,x)|\}$ for $i = 1,2$. If $\frac{\ln\rho}{T_a} + \beta < 0$, then the synchronization of the drive system (2) and the response system (3) is achieved.*

Remark 1 The idea of the proof for Theorem 1 is as follows: choose the Lyapunov function $V(t) = \frac{1}{2}\|\mathbf{e}(t,x)\|_2^2$; based on the idea of the proof for Theorem 2 in [2], show that $V'(t) \leq \beta V(t)$ on each impulsive interval. For $t = t_k$, we can follow the idea in [5] to get $V(t_k^+) \leq \rho V(t_k)$. Therefore, we can have $V(t) \leq V(t_0)\rho^{-N_0}e^{(\frac{\ln\rho}{T_a}+\beta)(t-t_0)}$, which implies that the synchronization can be realized exponentially with the convergence rate $-\frac{1}{2}(\frac{\ln\rho}{T_a} + \beta)$.

Remark 2 Based on Lyapunov function method, the impulsive synchronization criterion has been established. Compared with the existing result in [2], there are two improvements of this criterion: we derived an upper bound for the average impulsive interval, which is less conservative than the criterion in [2], since there is no strict restriction on the upper bound of each impulsive interval; pinning impulsive controller is designed which is added to one state of the Gray–Scott model at each impulsive instant. Let T_a be the upper bound of each impulsive interval and control all the states of the system (3) at each time; then Theorem 1 will reduce to a special case of the result in [2].

Remark 3 The criterion presented in Theorem 1 is closely related to the system parameters, the average impulsive interval T_a and the impulsive strength q. From Theorem 1, we can get the upper bound for the average impulsive interval:

$$T_a < \frac{1}{\beta} \ln (1 - q(q-2)/2). \tag{6}$$

However, the criterion in Theorem 1 is a sufficient condition, which means that the synchronization of the drive system (2) and the system (3) can be realized even if (6) does not hold.

Remark 4 The chaotic behavior of the Gray–Scott model guarantees the existence of β_i, that is, boundedness of $|u_i|$ and $|v_i|$. Therefore, even with slight perturbations in system parameters, if the system exhibits chaotic behaviors, then our impulsive control approach is still applicable to achieve the synchronization of the chaotic system.

3 Impulsive Stabilization of Spatiotemporal Chaos

Since $E_0 = (1,0)^T$ is a trivial state of the Gray–Scott model, if we choose $\mathbf{u}_0(x) = \mathbf{h}(t) = (1,0)^T$, then, from the system (2), $(u_1(t,x), u_2(t,x))^T \equiv (1,0)^T$.

Therefore, the synchronization problem of the drive system (2) and the response system (3) reduces to the stability problem of the equilibrium E_0 of the following impulsive partial differential system:

$$\begin{cases} \frac{\partial v_1}{\partial t} = -v_1 v_2^2 + a(1 - v_1) + d_1 \frac{\partial^2 v_1}{\partial x^2}, \ t \neq t_k, \\ \frac{\partial v_2}{\partial t} = v_1 v_2^2 - (a+b) v_2 + d_2 \frac{\partial^2 v_2}{\partial x^2}, \ t \neq t_k, \\ \Delta \mathbf{v}(t,x) = I_k(\mathbf{v}(t,x)), \ \ t = t_k, \ x \in [0, L], \ k = 1, 2, ..., \\ \mathbf{v}(0,x) = \mathbf{v}_0(x), \ \ x \in [0, L], \\ \mathbf{v}(t,0) = \mathbf{v}(t,L) = \mathbf{h}, \ \ t \in \mathbb{R}^+, \end{cases} \tag{7}$$

where $\mathbf{h} = (h_1, h_2)^T = (1,0)^T$.

The impulsive controller is designed as follows:

$$I_k(\mathbf{v}(t_k, x)) = \begin{cases} -q(h_i - v_i(t_k, x)), \ \ i = \mathcal{D}_k, \\ 0, \quad i \neq \mathcal{D}_k, \end{cases} \tag{8}$$

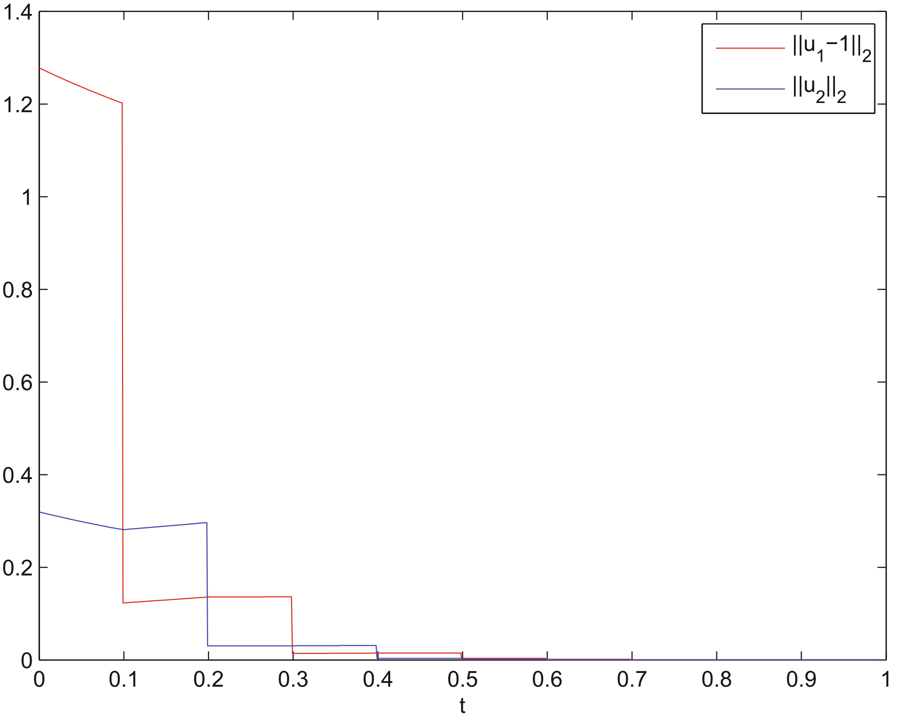


Fig. 2 Uniform impulsive intervals

where the index $\mathcal{D}_k$ is defined the same as in the controller (5) with $\mathbf{e}(t, x) = \mathbf{h} - \mathbf{v}(t, x)$.

We have the following stability result about the impulsive system (7), the proof of which is similar to the proof of Theorem 1.

Theorem 2 *Suppose the average impulsive interval of the impulsive sequence* $\zeta = \{t_k\}$ *is less than* T_a. *Let* $\rho = 1 - q(q-2)/2$ *and* $\beta = 4\beta_2\sqrt{\beta_1^2 + 4\beta_2^2} - 2a$, *where* $\beta_1 := \max\{\sup_{t\in\mathbb{R}^+} |v_1(t, x)|, 1\}$ *and* $\beta_2 := \sup_{t\in\mathbb{R}^+} |v_2(t, x)|$. *If* $\frac{\ln\rho}{T_a} + \beta < 0$, *then the equilibrium* E_0 *of the impulsive system (7) is globally asymptotically stable.*

Remark 5 From Theorem 2, we see that based on the Lyapunov function method the states of the one-dimensional Gray–Scott model are driven to the equilibrium $E_0 = (1, 0)^T$ effectively by a pinning impulsive controller. Actually, form Remark 1, we can see that the equilibrium E_0 of the impulsive system (7) is globally exponentially stable with the convergence rate $-\frac{1}{2}(\frac{\ln\rho}{T_a} + \beta)$. In the following numerical simulations, we choose $T_a = 0.1$ and $q = 0.78$, which implies that all the conditions of Theorem 2 are satisfied. Uniform impulsive intervals are chosen in Fig. 2 with $t_{k+1} - t_k = 0.1$, while $t_{2k} - t_{2k-1} = 0.04$ and $t_{2k+1} - t_{2k} = 0.16 > T_a$ are chosen in Fig. 3. We can see from Figs. 2 and 3 that the equilibrium E_0 of system (7) is asymptotically

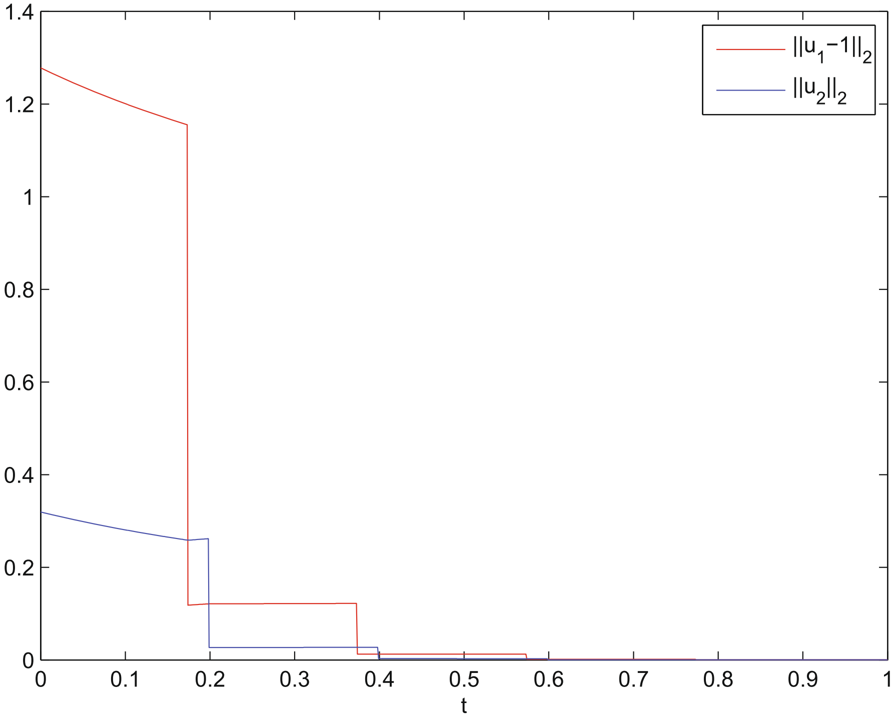


Fig. 3 Nonuniform impulsive intervals

stable. The numerical method used for Figs. 1–3 is the forward Euler integrations of the finite-difference equations.

References

1. Gray, P., Scott, S.K.: Autocatalytic reactions in the isothermal, continuous stirred reactor. Chem. Eng. Sci. **39**(6), 1087–1097 (1984)
2. Khadra, A., Liu, X., Shen, X.: Impulsive control and synchronization of spatiotemporal chaos. Chaos Solitions Fractals **26**(2), 615–636 (2005)
3. Kyrychko, Y.N., Blyuss, K.B.: Control of spatio-temporal patterns in the Gray–Scott model. Chaos. **19**(4), 043126 (2009)
4. Lu, J., Ho, D., Cao, J.: A unified synchronization criterion for impulsive dynamic networks. Automatica **46**(7), 1215–1221 (2010)
5. Lu, J., Kurths, J., Cao, J., Mahdavi, N., Huang, C.: Synchronization control for nonlinear stochastic dynamic networks: pinning impulsive strategy. IEEE Trans. Neural Netw. Learn. Syst. **23**(2), 285–292 (2012)
6. Nishiura, Y., Ueyama, D.: Spatio-temporal chaos for the Gray-scott model. Physica D **150**(3–4), 137–162 (2001)