

Springer Proceedings in Mathematics & Statistics

D. Marc Kilgour · Herb Kunze ·
Roman Makarov · Roderick Melnik ·
Xu Wang *Editors*

Recent Developments in Mathematical, Statistical and Computational Sciences

The V AMMCS International Conference,
Waterloo, Canada, August 18–23, 2019

 Springer

Springer Proceedings in Mathematics & Statistics

Volume 343

This book series features volumes composed of selected 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>

D. Marc Kilgour · Herb Kunze · Roman Makarov ·
Roderick Melnik · Xu Wang
Editors

Recent Developments in Mathematical, Statistical and Computational Sciences

The V AMMCS International Conference,
Waterloo, Canada, August 18–23, 2019

 Springer

Editors

D. Marc Kilgour
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON, Canada

Herb Kunze
Department of Mathematics and Statistics
University of Guelph
Guelph, ON, Canada

Roman Makarov
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON, Canada

Roderick Melnik
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON, Canada

Xu Wang
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON, Canada

ISSN 2194-1009

ISSN 2194-1017 (electronic)

Springer Proceedings in Mathematics & Statistics

ISBN 978-3-030-63590-9

ISBN 978-3-030-63591-6 (eBook)

<https://doi.org/10.1007/978-3-030-63591-6>

Mathematics Subject Classification: 15-XX, 34-XX, 35-XX, 37-XX, 60-XX, 45-XX, 49-XX, 62-XX, 65-XX, 68-XX, 74-XX, 76-XX, 91-XX, 92-XX

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains a selection of papers originally prepared for presentations at AMMCS-2019, an international conference held in Waterloo, Ontario, Canada from 18 to 23 August 2019. It was the fifth in the series of AMMCS meetings, held biennially at Wilfrid Laurier University beginning in 2011. The 2019 event continued the tradition of promoting interdisciplinary research and collaboration involving mathematical, statistical and computational sciences within the broader international community, highlighting recent advances in Applied Mathematics, Modeling and Computational Science (AMMCS).

The interdisciplinary focus of the AMMCS conferences has been crucial to their success. The primary aim of AMMCS-2019 was to promote research and collaboration involving new applications of mathematical, statistical and computational sciences to many fields, for the benefit of international communities of researchers, practitioners, and students.

For millennia, mathematical methods have been fundamental tools for the development of human knowledge. Now sophisticated mathematical and statistical tools are making essential contributions to progress in an amazing range of application areas—in the natural and social sciences, engineering, finance, and even the arts. Mathematics, statistics, and associated computational and data science techniques are playing a fundamental role in the modern world, throwing new light on problems, both ancient and contemporary, thereby contributing to human well-being.

Today's most challenging problems arise not only in the physical sciences and engineering, where mathematics is traditionally applied, but also in the life sciences, the social sciences, and finance. Stunning advances in these areas have resulted from the great subtlety and power of mathematical techniques and reasoning, augmented by data collection and analysis on a scale more massive than could be imagined only a few years ago, and by computational studies that not only support analysis but also explore new combinations and structures. These developments have forged new connections among disciplines that were once widely separated, as the horizons of mathematical and computational modeling expand at an increasing rate.

AMMCS-2019 was a major international forum for the exchange of ideas in an interdisciplinary setting, with a focus on applications of mathematical and

computational sciences, modeling and simulation to the natural and social sciences, engineering and technology, industry and finance. It proudly followed the traditions of previous AMMCS events, particularly in its emphasis on discussion, comparison, and synthesis across disciplines. We believe that only through interdisciplinary collaboration will it be possible to meet the complex challenges facing humanity today.

This book consists of a representative selection of current research presented at AMMCS-2019. It illustrates how mathematics, statistics, and modeling are contributing to a range of disciplines. The 68 selected contributions are organized into six parts, as follows:

- I. Advances in Mathematical Modelling and Theory;
- II. Advances in Statistical Modelling and Data Analysis;
- III. Computational Methods for Differential Equations;
- IV. Mathematical Modelling in Engineering, Physical and Chemical Sciences;
- V. Mathematical and Statistical Modelling in Life Sciences;
- VI. Mathematics and Computation in Finance, Economics, and Social Sciences.

The titles of the parts make the breadth of the topics clear. This wide-ranging selection shows clearly how mathematical, statistical, and computational sciences are now emerging as fundamental tools in a wide range of disciplines.

The editors of this volume extend their thanks to all of the contributors to AMMCS-2019, to all of the attendees, to the Organizing, Scientific, and Technical Committees, and to all of the volunteers, without whom the conference could not have taken place. We are grateful to our sponsors and to Wilfrid Laurier University. We give special thanks to the contributors who prepared papers for this volume, and to the referees whose guidance was essential to us as we evaluated proposed contributions. We also thank Leonie Kunz of Springer, who assisted us with the documents and editorial work, and Banu Dhayalan of Springer, who handled the technical aspects of production and publishing. We are proud of this volume, and pleased to acknowledge all those who helped bring it to fruition.



Group photo of participants in the AMMCS-2019 International Conference: Applied Mathematics, Modeling and Computational Science

Waterloo, ON, Canada
October 2020

D. Marc Kilgour
Herb Kunze
Roman Makarov
Roderick Melnik
Xu Wang

Contents

Advances in Mathematical Modelling and Theory

A Moving Horizon State Estimator for Real-Time Stabilization of a Double Inverted Pendulum	3
Amanda Bernstein, Ethan King, and Hien Tran	
Spontaneous Symmetry-Breaking in Deterministic Queueing Models with Delayed Information	15
Juancho A. Collera	
Algebraic Structure of the Varikon Box	27
Jason d'Eon and Chrystopher L. Nehaniv	
A Bestiary of Transformation Semigroups for the Holonomy Decomposition	37
Attila Egri-Nagy and Chrystopher L. Nehaniv	
Spatial Iterated Prisoner's Dilemma as a Transformation Semigroup	47
Isaiah Farahbakhsh and Chrystopher L. Nehaniv	
Oscillations and Periodic Solutions in a Two-Dimensional Differential Delay Model	59
Anatoli F. Ivanov and Zari A. Dzalilov	
Exploring Tetris as a Transformation Semigroup	71
Peter C. Jentsch and Chrystopher L. Nehaniv	
Differential Equations Using Generalized Derivatives on Fractals	81
Herb Kunze, Davide La Torre, Franklin Mendivil, and Edward R. Vrscay	
Revisiting Path-Following to Solve the Generalized Nash Equilibrium Problem	93
Tangi Migot and Monica-G. Cojocaru	

Properties of the Zeros of the Scale-Delay Equation and Its Time-Variant ODE Realization 103
 Erik I. Verriest

Advances in Statistical Modelling and Data Analysis

Covering Large Complex Networks by Cliques—A Sparse Matrix Approach 117
 W. M. Abdullah, S. Hossain, and M. A. Khan

Comparing Regularization Techniques Applied to a Perceptron 129
 Bryson Boreland, Herb Kunze, and Kimberly M. Levere

Key Performance Indicators and Individual Factors on Penalty Kicks 137
 Joao Fialho

Sparse Covariance and Precision Random Design Regression 147
 Xi Fang, Steven Winter, and Adam B. Kashlak

Applying Neural Networks to a Fractal Inverse Problem 157
 Liam Graham and Matthew Demers

Solving Parameter Identification Problems using the Collage Distance and Entropy 167
 Herb Kunze and Davide La Torre

Mean-Square Stability of Stochastic System with Impulse and Unbounded Delay 177
 Mengling Li, Feiqi Deng, and Xinzhi Liu

BOLD.R: A Software Package to Interface with BOLD Through R 187
 Nishan Mudalige

Analysis of Cortical Spreading Depression in Brain with Multiscale Mathematical Models 197
 Hina Shaheen, Roderick Melnik, and Sundeep Singh

Impulsive Consensus of Complex-Valued Multi-agent Systems with Hybrid Protocols 209
 Yuan Shen, Xianguo Li, and Xinzhi Liu

Input-to-State Stability for Delayed Hybrid Systems and H_∞ Control 221
 Taghreed G. Sugati, Mohamad S. Alwan, and Xinzhi Liu

Impulsive Distance-Based Formation Tracking Control of Multi-agent Systems 233
 Zixing Wu, Xinzhi Liu, and Jinsheng Sun

Exponential Stabilization for Markov Jump Neural Networks with Additive Time-Varying Delays via Event-Triggered Impulsive Control 243
 Haiyang Zhang, Zhipeng Qiu, Xinzhi Liu, and Lianglin Xiong

Computational Methods for Differential Equations

Development of a Lattice Boltzmann Model for the Solution of Partial Differential Equations, A Performance Comparison Study with that of the Finite Difference Method 255
 Mahmud Ashrafizaadeh and A. Ghavaminia

Using Shooting Approaches to Generate Initial Guesses for ODE Parameter Estimation 267
 Jonathan Calver, Jienan Yao, and Wayne Enright

A Computational Study for Solving Inverse Problems for Mixed Variational Equations on Perforated Domains 277
 A. I. Garralda-Guillem, Herb Kunze, Davide La Torre, and M. Ruiz Galán

bacoli_py—A Python Package for the Error Controlled Numerical Solution of 1D Time-Dependent PDEs 289
 Connor Tannahill and Paul Muir

Solving Cardiac Bidomain Problems with B-spline Adaptive Collocation 301
 Kevin R. Green and Raymond J. Spiteri

A Computational Comparison of Three Methods for Solving a 1D Boundary Value Inverse Problem 313
 Kimberly M. Levere, Bryson Boreland, and John Dewhurst

A Comparison of Turbulence Generated by 3DS Sparse Grids with Different Blockage Ratios and Different Co-frame Arrangements 325
 M. Syed Usama and Nadeem A. Malik

Mathematical Modelling in Engineering, Physical and Chemical Sciences

An Extended Pseudo Potential Multiphase Lattice Boltzmann Model with Variable Viscosity Ratio 337
 Mahmud Ashrafizaadeh, Farshad Gharibi, and Seyyed Meysam Khatoonabadi

Approximating Dispersive Materials with Parameter Distributions in the Lorentz Model 349
 Jacqueline Alvarez, Andrew Fisher, and Nathan L. Gibson

Coulomb Explosion Imaging: Super-Resolution by Optical Properties of Electrostatics Lenses	361
David Babalola and C. Sean Bohun	
Error Correction for Correlated Quantum Systems	373
Mark Byrd, Alvin Gonzales, Daniel Dilley, and Purva Thakre	
Numerical Investigation of VAWT Airfoil Shapes on Power Extraction and Self-starting Purposes	383
Sajad Maleki Dastjerdi, Amir HormoziNejad, Kobra Gharali, and Jatin Nathwani	
Stability and Stabilization of T–S Fuzzy Systems with Aperiodic Sampling	393
Jinnan Luo, Xinzhi Liu, Wenhong Tian, Shouming Zhong, and Kaibo Shi	
A New Method of Modelling Tuneable Lasers with Functional Composition	401
B. Metherall and C. Sean Bohun	
Algebraic Structure and Complexity of Bootstrap Percolation with External Inputs	411
S. Pal and Chrystopher L. Nehaniv	
Simulations of Realistic Trombone Notes in the Time-Domain	423
Janelle Resch, Lilia Krivodonova, and John Vanderkooy	
The Impact of External Features on Prediction Accuracy in Short-Term Energy Forecasting	431
Maher Selim, Ryan Zhou, Wenying Feng, and Omar Alam	
Toral Diffeomorphisms Induce Quantum Superoperators via TAQS	441
Artur Sowa	
Optimal Time Decay Rates for a Chemotaxis Model with Logarithmic Sensitivity	451
Yanni Zeng and Kun Zhao	
Mathematical and Statistical Modelling in Life Sciences	
An Optimal Control Strategy for a Malaria Model	465
Onoja Abu and Ikechukwu Ignatius Ayogu	
Effect of Genetic Defects in a Cortical Circuit Model Associated with Childhood Absence Epilepsy	477
Maliha Ahmed and Sue Ann Campbell	
Operator Splitting for the Simulation of Aqueous Humor Thermo-Fluid-Dynamics in the Anterior Chamber	489
Farah Abdelhafid, Giovanna Guidoboni, Naoki Okumura, Noriko Koizumi, and Sangly P. Srinivas	

Modelling Thermal Aspects of Decomposition	501
L. Calla and C. Sean Bohun	
Mathematical Modeling of the Steady-State Behavior of Nitric Oxide in Brain	511
Corina S. Drapaca and Andrew Tamis	
Automate Obstructive Sleep Apnea Diagnosis Using Convolutional Neural Networks	521
Longlong Feng and Xu Wang	
Numerical Analysis of Nanowire Resonators for Ultra-high Resolution Mass Sensing in Biomedical Applications	533
Rosa Fallahpour and Roderick Melnik	
Contaminant Removal in Ceramic Water Filters by Bacterial Biofilms	545
Harry J. Gaebler, Jack M. Hughes, and Hermann J. Eberl	
Comparison of Fractional-Order and Integer-Order Cancer Tumor Growth Models: An Inverse Approach	555
Jennifer Lawson and Kimberly M. Levere	
Numerical Modelling of Drug Delivery in an Isolated Solid Tumor Under the Influence of Vascular Normalization	565
Mahya Mohammadi, Cyrus Aghanajafi, and Madjid Soltani	
Quantitative Study of the Coupling Among Cardiovascular System, Lymphatic System and Interstitial Space	579
Nicholas Mattia Marazzi, Virginia H. Huxley, Riccardo Sacco, and Giovanna Guidoboni	
Age-Structured Epidemic with Adaptive Vaccination Strategy: Scalar-Renewal Equation Approach	591
Aubain Nzokem and Neal Madras	
On the Modeling of Drug Delivery to Solid Tumors; Computational Viewpoint	601
Mohsen Rezaeian, Madjid Soltani, and Farshad Moradi Kashkooli	
Evaluating a Logistic K-mer Based Model for Classifying CO1 Sequences of <i>C. Clupeaformis</i>	611
D. St Jean, Herb Kunze, and D. Gillis	
Mathematical Modeling of Coupled Electro-thermal Response of Nerve Tissues Subjected to Radiofrequency Fields	621
Sundeeep Singh and Roderick Melnik	

Ranking Association Rules from Data Mining for Health Outcomes: A Case Study of Effect of Industrial Airborne Pollutant Mixtures on Birth Outcomes 633
 K. Vu, A. Osornio-Vargas, O. Zaiane, and Y. Yuan

Mathematics and Computation in Finance, Economics, and Social Sciences

About the Algorithms of Strategic Management 647
 Manana Chumberidze, Mzia Kiknadze, Nino Topuria, and Elza Bitsadze

Using Cognitive Fit Theory to Evaluate the Effectiveness of Financial Information Visualization: An Example Using Data to Detect Fraudulent Transactions 655
 A. Czeglédi, L. Scott Campbell, and D. Smiderle

Factors Affecting Sustainable Development and Modelling 669
 Zurab Gasitashvili, Mzia Kiknadze, Taliko Zhvania, and David Kapanadze

On a Generalized Integro-Differential Spatial Model of Economic Growth 681
 Herb Kunze, Davide La Torre, and Simone Marsiglio

Utilizing Bidirectional Encoder Representations from Transformers for Answer Selection 693
 Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang

Calibration and Analysis of Structural Credit Risk Models with Occupation Time 705
 Malhar M. Mukhopadhyay and Roman N. Makarov

Prediction Intervals of Machine Learning Models for Taxi Trip Length 715
 Ella Morgan, Ryan Zhou, and Wenying Feng

The Cobb-Douglas Production Function Revisited 725
 Roman G. Smirnov and Kunpeng Wang

Inferring Rankings from First Order Marginals 735
 Sarah Wolff

High-Frequency Statistical Modelling for Jump-Diffusion Multi-asset Price Processes with a Systemic Component 747
 Rulin Xu and Roman N. Makarov

Advances in Mathematical Modelling and Theory

A Moving Horizon State Estimator for Real-Time Stabilization of a Double Inverted Pendulum



Amanda Bernstein, Ethan King, and Hien Tran

Abstract A moving horizon estimator is designed in the framework of the proximal point minimization algorithm for linear time invariant systems and convergence results are established in the presence of model and measurement noise. The state estimator is implemented in a nonlinear suboptimal feedback control framework for the real time stabilization of a double inverted pendulum on a cart.

Keywords Time invariant systems · Measurement noise · State estimations · Minimization algorithms · Nonlinear feedback stabilization

1 Introduction

The double inverted pendulum (DIP) is commonly used as a benchmark problem in nonlinear control theory as it is an under-actuated system with highly nonlinear dynamics which is amenable to laboratory study. Control of the DIP can also provide a direct model for systems such as robotic limbs [15], human posture, balance, and gymnast motion [17, 19].

Many control designs have been applied to the DIP system including the linear quadratic regulator (LQR) [7, 9], state dependent Riccati equation control [7], and neural network control [7, 18]. Less work has been done for state estimation of the

A. Bernstein · E. King (✉) · H. Tran
Department of Mathematics, Center for Research in Scientific Computation,
North Carolina State University, Raleigh, NC 27695, USA
e-mail: eaking4@ncsu.edu

H. Tran
e-mail: tran@ncsu.edu

Present Address:
A. Bernstein
U.S. EPA, Durham, NC 27709, USA
e-mail: ascoons@ncsu.edu

DIP as many studies use only simulations. State estimation though, is necessary for real-time experimental set ups and has been accomplished with Luenburger type observers [11] and low pass derivative filters [6].

The DIP on a cart, consists of two pendula in tandem connected on a hinge to a cart which moves on a linear track as shown in Fig. 1. Stabilization control refers to moving the cart along the track, such that the pendula are balanced vertically over the cart in an upright unstable equilibrium. While the pendulum angles and cart position are directly measured, for feedback control of the system, the velocity of the cart and pendulum angles must be estimated.

State estimation for DIP stabilization control has been found to be a challenging task for some methods [5]. The stabilization control of the DIP from an upright start provides an interesting problem, as it can require rapid convergence of state estimates, from a poor initial estimate, in order for the feedback control to rescue the system. The estimation method must also achieve robust performance in the presence of significant model error, measurement error, and disturbances to be effective.

This paper studies state estimation and stabilization control of a DIP on a cart. A nonlinear feedback stabilization control is designed using power series expansion following Garrard [10]. State estimation is approached from a proximal point perspective to construct a moving horizon estimator.

Moving horizon estimation (MHE) most commonly minimizes a least squares cost functional, which includes a regularizing term frequently called the arrival cost, to fit a model trajectory to the N most recent system measurement outputs. Inclusion of an arrival cost term has been found to be important to ensure convergence and stability properties of MHE algorithms [2, 13]. The arrival cost penalizes the difference between the new and previous state estimate and has been interpreted in several ways; as an estimate of the error of the fit to the truncated measurement history when MHE is viewed as approximating the Kalman filter [13], approximating use of an a-priori distribution in probabilistic settings [8], more loosely as a confidence in the past estimate [2], and can also be viewed as a regularization term for solving the state to output inverse problem when MHE is considered as an approximation of a deadbeat observer.

MHE algorithms have been found to converge quickly from poor initial estimates [3, 12, 13] and to perform robustly in the presence of noise [1, 3]. Similar results have been observed in online applications, including charge estimation of batteries [16] and state estimation of a vibrating active cantilever [1].

We approach iterative minimization for state estimation through the framework of the proximal operator and proximal point minimization algorithm. The proximal point minimization algorithm as given in [14] naturally gives rise to a quadratic regulating term similar to the arrival cost terms shown to be effective for MHE in practice. We develop and show convergence of a linear discrete time state estimator in this framework, similar to the MHE given by Alessandri et al. in [2], and implement it for stabilization of the DIP system.

This paper is organized as follows. In Sect. 2, we construct a proximal point MHE state estimator and give convergence results. In Sect. 3, we introduce the double inverted pendulum system and mathematical model. Section 4 describes a power

series based feedback stabilization control, and we implement the control and state estimate on the physical DIP system in Sect. 5.

2 A Proximal Point Moving Horizon Estimator for Linear Time Invariant Systems

Let $\Phi \in \mathbb{R}^{d \times d}$, $W \in \mathbb{R}^{d \times v}$, $C \in \mathbb{R}^{m \times d}$, and controls $u \in \mathbb{R}^v$. Consider the following discrete dynamical system

$$\begin{aligned} x_{k+1} &= \Phi x_k + W u_k + \eta_k \\ y_{k+1} &= C x_{k+1} + \varepsilon_k, \end{aligned} \quad (1)$$

where $\{\eta_k\}_{k \in \mathbb{N}}$ and $\{\varepsilon_k\}_{k \in \mathbb{N}}$ are unknown model and measurement noise, respectively.

Given the past N system outputs $\{y_{k-i}\}_{i=N-1}^0$, an estimate for the state at x_k is constructed with a moving horizon type state estimator using a least squares cost functional, by solving a problem of the form

$$\begin{aligned} \min_{\{z_{k-j}\}_{j=N-1}^0} & \|z_{k-(N-1)} - \hat{x}_{k-(N-1)}\|^2 + \sum_{i=N-1}^0 \|y_{k-i} - C z_{k-i}\|^2 \\ \text{subject to} & \quad z_{k-i+1} = \Phi z_{k-i} \quad \forall i \in 0, 1, \dots, N-1, \end{aligned} \quad (2)$$

where $\hat{x}_{k-(N-1)}$ is the previous state trajectory estimate and the first term in the cost functional (2) is taken as the arrival cost type term.

We approach the minimization at each iteration of MHE as applying the proximity operator of the trajectory fitting functional to the previous state estimate.

Definition 1 Let the function $\phi : \mathbb{R}^d \rightarrow [-\infty, \infty]$ and $\gamma \in]0, \infty[$. The proximity operator of $\gamma\phi$ is defined by

$$Prox_{\gamma\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^d : x \rightarrow \operatorname{argmin}_{z \in \mathbb{R}^d} \phi(z) + \frac{1}{2} \frac{1}{\gamma} \|z - x\|^2.$$

Iterative application of the proximity operator can generate a minimizing sequence for a given cost functional. In particular if ϕ convex, proper, and lower semi-continuous, with $\operatorname{argmin} \phi \neq \{\emptyset\}$ then for all $\gamma \in]0, \infty[$, $Prox_{\gamma\phi}$ is well defined. Moreover, for any initial value $z_0 \in \mathbb{R}^d$ with sequence $\{\gamma_k\}_{k \in \mathbb{N}}$ in $]0, \infty[$ such that $\sum_{k \in \mathbb{N}} \gamma_k = \infty$ the proximal point iteration

$$z_{k+1} = Prox_{\gamma_k \phi} z_k$$

generates a minimizing sequence of ϕ [4].

We consider the design of a discrete time state estimator for the system (1) by constructing functionals $\{\phi_k\}_{k \in \mathbb{N}}$ such that $\operatorname{argmin}_{z \in \mathbb{R}^d} \phi_k(z) = x_k$. Then we define the

proximal point observer as the sequences $\{\hat{x}_k\}_{k \in \mathbb{N}}$ and $\{p_k\}_{k \in \mathbb{N}}$ constructed according to the recursion

$$\begin{aligned} p_k &= \text{Prox}_{\gamma_k \phi_k} \hat{x}_k \\ \hat{x}_{k+1} &= f(p_k). \end{aligned} \quad (3)$$

Functionals capturing least squares type moving horizon estimators can be constructed with a $G \in \mathbb{R}^{(N-m) \times d}$ giving the model to output map and vector $v_k \in \mathbb{R}^{N-m}$ incorporating the model and measurement noise, in the form

$$\phi_k(z) \doteq \frac{1}{2} \|G(x_k - z) + v_k\|^2. \quad (4)$$

For example for the MHE given in (2)

$$G = \begin{bmatrix} C\Phi \\ \dots \\ C(\Phi)^N \end{bmatrix}, \quad v_k = \begin{bmatrix} C\eta_k + \varepsilon_k \\ \dots \\ C \sum_{j=0}^{N-2} \Phi^{N-1-j} \eta_{k+j} + C\eta_{(k+(N-1))} + \varepsilon_{k+N} \end{bmatrix}.$$

Suppose that the matrix $G^T G$ is positive definite, then each ϕ_k has a unique minimizer z_k^* , given by

$$z_k^* = x_k + (G^T G)^{-1} G^T v_k.$$

The discrepancy between the minimizer and the true state of the system is denoted as the noise term ζ_k , where for each functional

$$\zeta_k = (G^T G)^{-1} G^T v_k.$$

For convenience the functionals are also adjusted such that the minimum value is zero as follows

$$\phi_k(z) = \frac{1}{2} \|G(x_k - z) + v_k\|^2 - \frac{1}{2} \|(I - G(G^T G)^{-1} G^T)v_k\|^2,$$

which may be written more conveniently as

$$\phi_k(z) = \frac{1}{2} \|G((x_k + \zeta_k) - z)\|^2. \quad (5)$$

Using functionals of the form (5) the error for the proximal observer estimates (3) follows a simple recursion. For the following it is assumed $G^T G$ is positive definite, with $U \in \mathbb{R}^{d \times d}$ a unitary matrix, and $\Lambda \in \mathbb{R}^{d \times d}$ diagonal such that $G^T G = U \Lambda U^T$. The eigenvalues of $G^T G$ are denoted by $\{\lambda_i\}_{i=1}^d$.

From an initial estimate \hat{x}_0 , let the sequences $\{\hat{x}_k\}_{k \in \mathbb{N}}$ and $\{p_k\}_{k \in \mathbb{N}}$ be generated according to (3) using the cost functionals (5), with sequence of weighting parameters $\{\gamma_k\}_{k \in \mathbb{N}}$ in $\mathbb{R}_{>0}$.

Proposition 1 *The error terms $e_k = (\hat{x}_k - x_k)$ satisfy the recursion*

$$e_{k+1} = \Phi U \bar{\Lambda}_k U^T e_k + \Phi U \ddot{\Lambda}_k U^T \zeta_k - \eta_k$$

where the diagonal matrices $\bar{\Lambda}_k$, $\ddot{\Lambda}_k$ have entries $\bar{\Lambda}_{i,i} = \frac{1}{1 + \gamma_k \lambda_i}$, and $\ddot{\Lambda}_{i,i} = \frac{\gamma_k \lambda_i}{1 + \gamma_k \lambda_i}$.

Proof Here we provide only the main steps of the proof. Note from (3)

$$p_k = \text{Prox}_{\gamma_k \phi_{G_k}}(\hat{x}_k) = \underset{z \in \mathbb{R}^n}{\text{argmin}} \left\{ \frac{1}{2} \|G(z - (x_k + \zeta_k))\|^2 + \frac{1}{2\gamma_k} \|z - \hat{x}_k\|^2 \right\}.$$

Then computing the gradient and setting it equal to zero yields

$$p_k = (G^T G + \frac{1}{\gamma_k} I)^{-1} G^T G(x_k + \zeta_k) + \frac{1}{\gamma_k} (G^T G + \frac{1}{\gamma_k} I)^{-1} \hat{x}_k.$$

Using the fact $G^T G = U \Lambda U^T$,

$$p_k = U(\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda U^T x_k + \frac{1}{\gamma_k} U(\Lambda + \frac{1}{\gamma_k} I)^{-1} U^T \hat{x}_k + U(\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda U^T \zeta_k.$$

Therefore,

$$(p_k - x_k) = U((\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda - I) U^T x_k + \frac{1}{\gamma_k} U(\Lambda + \frac{1}{\gamma_k} I)^{-1} U^T \hat{x}_k + U(\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda U^T \zeta_k.$$

Note that

$$(\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda - I = -\bar{\Lambda}_k, \quad \frac{1}{\gamma_k} (\Lambda + \frac{1}{\gamma_k} I)^{-1} = \bar{\Lambda}_k, \quad (\Lambda + \frac{1}{\gamma_k} I)^{-1} \Lambda = \ddot{\Lambda}_k$$

then

$$(p_k - x_k) = U \bar{\Lambda}_k U^T e_k + U \ddot{\Lambda}_k U^T \zeta_k.$$

Therefore,

$$e_{k+1} = (\hat{x}_{k+1} - x_{k+1}) = \Phi(p_k - x_k) - \eta_k = \Phi U \bar{\Lambda}_k U^T e_k + \Phi U \ddot{\Lambda}_k U^T \zeta_k - \eta_k.$$

The error recursion of Proposition 1, can be used to choose weighting parameters $\{\gamma_k\}_{k \in \mathbb{N}}$ to ensure the error is small relative to the noise. In particular, suppose for all $k \in \mathbb{N}$, $\gamma_k = \gamma$, then the proximal points p_k can be computed more efficiently at each iteration. A criteria for selecting a fixed γ follows immediately from Proposition 1.

Proposition 2 *If λ_{\min} the smallest eigenvalue of $G^T G$ and $\gamma > \max \left\{ \frac{\|\Phi\| - 1}{\lambda_{\min}}, 0 \right\}$, then*

$$\lim_{k \rightarrow \infty} \|e_k\| \leq \frac{\kappa}{1 - r},$$

where $\kappa = \|\Phi\| \bar{\zeta} + \bar{\eta}$ and $r = \frac{\|\Phi\|}{1 + \gamma \lambda_{\min}}$.

Proof Using proposition 1 for all $k \in \mathbb{N}$

$$\begin{aligned} \|e_{k+1}\| &\leq \|\Phi\| \frac{1}{1 + \gamma \lambda_{\min}} \|e_k\| + \|\Phi\| \frac{\gamma \lambda_{\max}}{1 + \gamma \lambda_{\max}} \|\zeta_k\| + \|\eta_k\| \\ &\leq r \|e_k\| + \kappa \end{aligned}$$

Therefore,

$$\|e_k\| \leq \|e_0\| r^k + \kappa \sum_{m=0}^{k-1} r^m,$$

and $r < 1$, hence

$$\lim_{k \rightarrow \infty} \|e_k\| \leq \frac{\kappa}{1 - r}.$$

2.1 Centered Proximal Point Observer

Many cost functionals of the form (5) can be constructed for discrete systems (1). A functional utilizing only the three most recent measurements was found to be effective for the DIP state estimation problem. For all $k \in \mathbb{N}$ let the function $\phi_{ctrk} : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as

$$\phi_{ctrk}(z) \doteq \frac{1}{2} \|C(\Phi^{-1}z + W^{-1}u_{k-1}) - y_{k-1}\|^2 + \frac{1}{2} \|Cz - y_k\|^2 + \frac{1}{2} \|C(\Phi z + Wu_k) - y_{k+1}\|^2. \quad (6)$$

Then with $v_k \in \mathbb{R}^{3 \cdot m}$ and $G \in \mathbb{R}^{(3 \cdot m) \times d}$ given by

$$G = \begin{bmatrix} C\Phi^{-1} \\ C \\ C\Phi \end{bmatrix} \quad \text{and} \quad v_k = \begin{bmatrix} -C\Phi^{-1}\eta_{k-1} + \varepsilon_{k-1} \\ \varepsilon_k \\ C\eta_k + \varepsilon_{k+1} \end{bmatrix},$$

ϕ_{ctrk} is of the form (5).

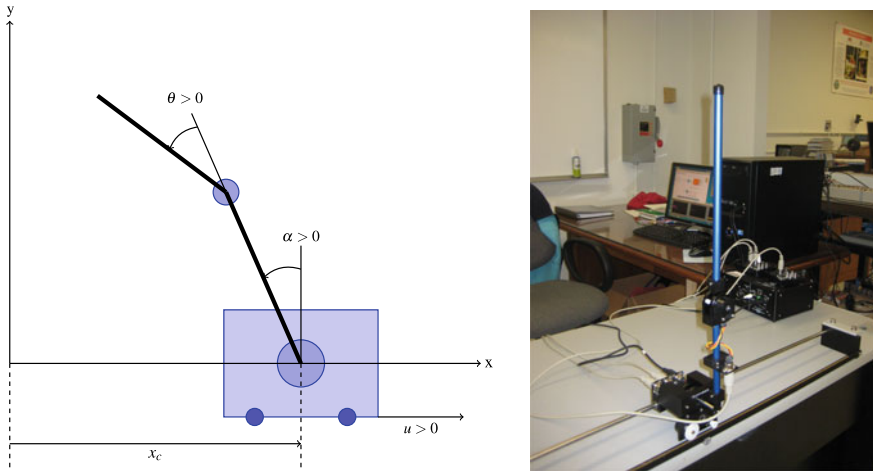


Fig. 1 Diagram and photo of the DIP system

3 Mathematical Model of the Double Inverted Pendulum

The DIP system used in this work was provided by Quanser Consulting Inc. and consists of an upper (12 in.) aluminium rod connected on a hinge to a lower (7 in.) rod which is in turn connected on a hinge to a cart (an IPO2 linear servo unit) that moves on a track as shown in Fig. 1. Encoders measure three variables; the position of the cart (x_c), the angle between the lower pendulum and normal vector vertical to the cart (α), and the angle between the lower and upper pendulum (θ). The measurements are defined such that upright unstable equilibrium is the origin and counterclockwise rotation is positive. The system is controlled through voltage input to a DC motor that moves the cart along the track.

The model for the DIP is derived using Lagrange's energy method as is commonly done, for example in [7, 9, 11]. A full derivation and description of the model can be found in [5, 6].

4 Power Series Stabilization Controller

For an $f : \mathbb{R}^6 \rightarrow \mathbb{R}^6$ and $B \in \mathbb{R}^6$, with control, $u : [0, \infty[\rightarrow \mathbb{R}$, the DIP dynamics are given by a system of the form

$$\dot{x} = f(x) + Bu \quad \text{for } x = [x_c, \theta, \alpha, \dot{x}_c, \dot{\theta}, \dot{\alpha}]^T.$$

A feedback stabilization control is constructed with respect to the cost functional

$$J(x_0, u) = \frac{1}{2} \int_0^\infty x^T Q x + R u^2 dt, \quad (7)$$

with $Q \in \mathbb{R}^{6 \times 6}$ and $R \in \mathbb{R}$, symmetric positive definite matrices. The optimal feedback control is

$$u(x) \doteq -R^{-1} B^T V_x(x),$$

where V is the solution to the corresponding Hamilton Jacobi Bellman equation, which is approximated using power series expansions following [10]. Let f be expanded about the unstable equilibrium as

$$f(x) = Ax + \sum_{n=2}^{\infty} f_n(x), \quad \text{where } f_n(x) = O(|x|^n).$$

Then for P the solution of the algebraic Riccati equation corresponding to the linearized system and cost functional (7), the control for the DIP is given in the feedback form

$$u^*(x) = -R^{-1} B^T \left[Px - (A^T - P B R^{-1} B^T)^{-1} P f_3(x) \right]. \quad (8)$$

Full details can be found in [5].

5 Real Time Stabilization of a Double Inverted Pendulum

The centered proximal point observer was applied to the DIP system by first constructing a discrete system of the form (1) using a linearization of the nonlinear DIP model about the unstable equilibrium. The state transition matrices Φ , Φ^{-1} , W , and W^{-1} for the linearized system were approximated using Matlab's `expm` command. A fixed ($\gamma > 0$) was used in the computation of state estimates according to proximal point observer (3) using the cost functionals (6), which at each iteration requires a solution to

$$p_k = \operatorname{argmin}_{z \in \mathbb{R}^6} \frac{1}{2} \|Hz - q_k\|^2, \quad (9)$$

for

$$H = \begin{bmatrix} \frac{1}{\gamma} I \\ C \Phi^{-1} \\ C \\ C \Phi \end{bmatrix} \quad \text{and} \quad q_k = \begin{bmatrix} \frac{1}{\gamma} \hat{x}_k \\ y_{k-1} - C W^{-1} B u_{n-1} \\ y_k \\ y_{k+1} - C W B u_n \end{bmatrix}.$$

To compute (9), an offline QR factorization for $H = Q_H R_H$ was computed with the Matlab `qr` command. Then the centered proximal point moving horizon estimation (CPX) with a fixed γ was iterated from initial estimate $\hat{x}_0 = 0$ according to

$$\begin{aligned} p_k &= R_H^{-1} Q_H^T q_k \\ \hat{x}_{k+1} &= \Phi p_k + W u_k. \end{aligned} \quad (10)$$

When implemented for the DIP feedback control in real time, the estimates supplied to compute the control were the model predictions $\tilde{x}_{k+2} = \Phi \hat{x}_{k+1} + u_{k+1}$.

The power series feedback stabilization control was computed with

$$Q = \text{diag}([80, 300, 100, 0, 0, 0]) \text{ and } R = 0.5,$$

where the state of the system is $x = [x_c, \theta, \alpha, \dot{x}_c, \dot{\theta}, \dot{\alpha}]^T$.

The estimator and control were implemented in real time through MATLAB Simulink interfaced with Quanser's Quarc software on a desktop computer running Windows 7 with a 3.20 GHz Intel Core i5 650 processor and 4 GB of RAM, connected to the DIP system by two Q2-USB DAQ control boards, with the control voltage applied to the cart by a VoltPAQ amplifier.

For a comparison study, both a CPX and a second order low pass derivative filter (LDF) were used to supply state estimates for stabilization control. The CPX estimator (10) was applied with $\gamma = 150$, while the LDF was used with Quanser's supplied parameters: cutoff frequency $\omega = 100\pi$ for the cart, $\omega = 20\pi$ for the pendulum angles, and damping ratios 0.9. Stabilization control was initiated once measurement values were brought to within 0.01 of the balanced state, the average value and variance for the measured DIP states when under stabilization control with CPX and LDF are reported in Table 1. Feedback control using the CPX estimates maintained the system closer to the balanced state and with less variance than with LDF estimates.

The CPX and LDF differed most for the estimates of the rate of change for θ , the angle between the pendulums. Figure 2 shows a comparison between CPX and LDF angle velocity estimates using measurement data from the physical DIP system under stabilization control.

The criteria for γ to guarantee CPX estimate convergence given in Proposition 2 is $\gamma \geq 24,400$ for the DIP system. When values of γ satisfying the condition were used the CPX angle velocity estimates had large amplitude high frequency oscillations unsuitable for computing a control, γ was reduced two orders of magnitude from the

Table 1 Output of DIP stabilization over (6.5 s) interval using either centered proximal point MHE (CPX) or low pass derivative filter (LDF) to compute the feedback control, the stabilized state is the origin

	x_c (cm)		α ($^\circ$)		θ ($^\circ$)	
	Mean	Variance	Mean	Variance	Mean	Variance
CPX	-0.002	5.52	0.178	35.2	-0.45	3.89
LDF	-0.119	5.98	1.21	41.8	-0.867	6.78

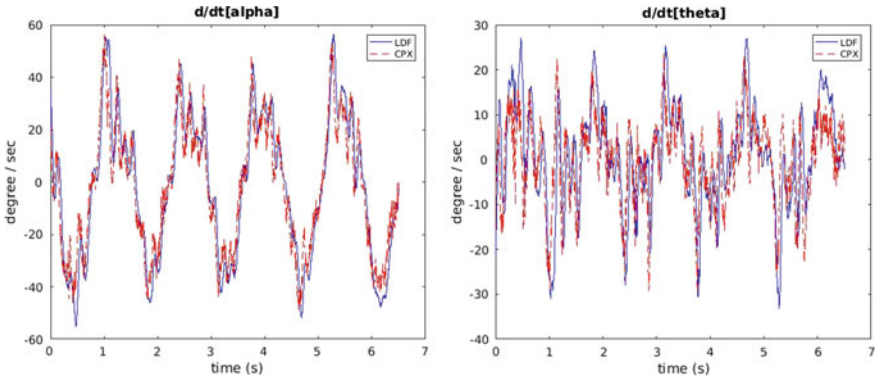


Fig. 2 Comparison of CPX and LDF angle velocity estimates for the real time DIP system under stabilization control

Proposition 2 criteria before a reasonable control could be computed. The need for a smaller γ value is likely due to H in (9) becoming more ill conditioned for larger γ and the solutions of (9) more sensitive to noise.

References

1. Abdollahpouri, M., Takács, G., Ilkiv, B.R.: Real-time moving horizon estimation for a vibrating active cantilever. *Mech. Syst. Signal Process.* **86**, 1–15 (2017)
2. Alessandri, A., Baglietto, M., Battistelli, G.: Receding-horizon estimation for discrete-time linear systems. *IEEE Trans. Autom. Control* **48**(3), 473–478 (2003)
3. Alessandri, A., Baglietto, M., Battistelli, G.: Moving-horizon state estimation for nonlinear discrete-time systems: new stability results and approximation schemes. *Automatica* **44**(7), 1753–1765 (2008)
4. Bauschke, H.H., Combettes, P.L.: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer (2017)
5. Bernstein, A.: Modeling and control: applications to a double inverted pendulum and radio frequency interference. Ph.D. Thesis, North Carolina State University, Raleigh, NC (2018)
6. Bernstein, A., Tran, H.T.: Real-time implementation of a LQR-based controller for the stabilization of a double inverted pendulum. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pp. 245–250 (2017)
7. Bogdanov, A.: Optimal control of double inverted pendulum on a cart. Oregon Health and Science University, Technical Report CSE-04-006, OGI School of Science and Engineering, Beaverton, OR (2004)
8. Delgado, R., Goodwin, G.C.: A combined MAP and Bayesian scheme for finite data and/or moving horizon estimation. *Automatica* **50**(4), 1116–1121 (2014)
9. Demirci, M.: Design of feedback controllers for linear system with applications to control of a double-inverted pendulum. *Int. J. Comput. Cogn.* **2**(1), 65–84 (2004)
10. Garrard, W.L.: Suboptimal feedback control for nonlinear systems. *Automatica* **8**, 219–221 (1972)
11. Graichen, K., Treuer, M., Zeitz, M.: Swing-up of the double pendulum on a cart by feedforward and feedback control with experimental validation. *Automatica* **43**(1), 63–71 (2007)

12. Kühn, P., Diehl, M., Kraus, T., Schlöder, J.P., Bock, H.: A real-time algorithm for moving horizon state and parameter estimation. *Comput. Chem. Eng.* **35**, 71–83 (2011)
13. Rao, C.V., Rawlings, J.B., Mayne, D.Q.: Constrained state estimation for nonlinear discrete-time systems: stability and moving horizon approximations. *IEEE Trans. Autom. Control* **48**(2), 246–258 (2003)
14. Rockafellar, R.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**(5), 877–898 (1976)
15. Saito, F., Fukuda, T., Arai, F.: Swing and locomotion control for two-link brachiation robot. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 719–724 (1993)
16. Shen, J., He, Y., Ma, Z., Luo, H., Zhang, Z.: Online state of charge estimation of lithium-ion batteries: a moving horizon estimation approach. *Chem. Eng. Sci.* **154**, 42–53 (2016)
17. Takashima, S.: Control of gymnast on a high bar. In: *Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems*, pp. 1424–1429 (1991)
18. Vicentini, F.: Stability analysis of evolved continuous time recurrent neural networks that balance a double inverted pendulum on a cart. In: *Proceedings of the International Joint Conference on Neural Networks*, pp. 2689–2694, Aug 2007
19. Winter, D.A.: Human balance and posture control during standing and walking. *Gait Posture* **3**(4), 193–214 (1995)

Spontaneous Symmetry-Breaking in Deterministic Queueing Models with Delayed Information



Juancho A. Collera

Abstract The dynamics of a system involving two queues which incorporates customer choice behavior based on delayed queue length information was studied recently. Waiting times in emergency rooms of hospitals, telephone call centers, and various rides in theme parks are some examples where delayed information is provided to the customers. This time delay has an impact on the dynamics of the queues and therefore has the capacity to affect the decision of a customer to choose which queue to wait in. We generalize this queueing model to a finite arbitrary number of queues. The system of delay differential equations for this generalized model is equivariant under a symmetry group. Spontaneous symmetry-breaking occurs in an equivariant dynamical system when the symmetry group of a solution of the equations is lesser than the symmetry group of the equations themselves. In this work, we show that the generalized model exhibits spontaneous symmetry-breaking. In particular, we show that varying the time delay parameter can make a stable equilibrium become unstable, and this switch in stability occurs only at a symmetry-breaking Hopf bifurcation. However, if the number of queues is chosen to be large enough, then the equilibrium is absolutely stable.

Keywords Delay differential equations · Queues · Delayed information · Spontaneous symmetry-breaking

1 Introduction

In queueing theory, a *fluid model* is a mathematical model which describes the fluid level in a reservoir where the periods of filling and emptying are randomly determined. More recently, fluid models are being used in nonlinear dynamics to describe certain applications such as in [7] which describes a single queue where the rate of

J. A. Collera (✉)

Department of Mathematics and Computer Science, University of the Philippines Baguio,
Governor Pack Road, Baguio 2600, Philippines
e-mail: jacollera@up.edu.ph

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_2

change with respect to time of a queue length is the difference between the arrival rate of the customers and the rate at which customers are serviced. In [8], the following deterministic queueing model where customers are given the opportunity to choose between two queues was introduced

$$\begin{cases} \frac{d}{dt}x(t) = a \frac{\exp(-x(t-\tau))}{\exp(-x(t-\tau)) + \exp(-y(t-\tau))} - bx(t), \\ \frac{d}{dt}y(t) = a \frac{\exp(-y(t-\tau))}{\exp(-x(t-\tau)) + \exp(-y(t-\tau))} - by(t). \end{cases} \quad (1)$$

Here, $x(t)$ and $y(t)$ are, respectively, the length of the first and second queues, and the time delay $\tau > 0$. For $t \in [-\tau, 0]$, the continuous functions $x(t) = \varphi_1(t) > 0$ and $y(t) = \varphi_2(t) > 0$ were used as initial history functions for system (1). It is assumed that the total arrival rate to both queues is equal to the constant rate a . Furthermore, since customers are given the queue length information τ time units in the past, the arrival rates in system (1) are based on *delayed information*.

We generalize the queues-with-choice model (1) introduced in [8] to an arbitrary but finite number of queues, and then study it from a symmetry perspective [5]. The generalized model is given by the following system of delay differential equations

$$\frac{d}{dt}x_k(t) = a \frac{\exp(-x_k(t-\tau))}{\sum_{n=1}^N \exp(-x_n(t-\tau))} - bx_k(t), \quad k = 1, 2, \dots, N. \quad (2)$$

The symmetry properties of system (2) play a significant role in determining its dynamical behavior. When a solution of the equations has a smaller symmetry group than the equations themselves, then we have a case called *spontaneous symmetry-breaking* [5]. This typically happens when a fully symmetric solution becomes unstable and branches of solutions with lesser symmetry bifurcate [11].

In this paper, we show that the symmetry structure of the generalized model (2) can be used to classify the types and kinds of solutions that can occur in the system. We show that spontaneous symmetry-breaking occurs in the system. In particular, this result explains why only asynchronous periodic solutions arise in such model, while synchronous periodic solutions do not. We also show that the number of queues can be chosen so that the equilibrium is *absolutely stable*, i.e. asymptotically stable for all time delays [1]. These additional insights on the dynamical behavior of queues will help managers of queues to be more aware of the consequences of providing delayed queue length information to their customers.

This paper is organized as follows. Section 2 provides preliminary materials needed to establish our results. Section 3 contains the results for an analysis of the generalized model from a symmetry perspective. The last section, which is Sect. 4, provides a summary and conclusion of the paper.

2 Preliminaries

In this section, we discuss what we mean by symmetries of a system of delay differential equations (DDEs). We consider the following class of DDEs

$$\frac{d}{dt}\mathbf{X}(t) = \mathbf{F}(\mathbf{X}(t), \mathbf{X}(t - \tau)) \quad (3)$$

with a single discrete time delay $\tau > 0$ and $\mathbf{X} : \mathbb{R} \rightarrow \mathbb{R}^N$ where N is a positive integer. We denote by $\mathcal{C} = C([-\tau, 0], \mathbb{R}^N)$ the space of continuous functions mapping the interval $[-\tau, 0]$ into \mathbb{R}^N . Together with the supremum norm, $\|\phi\|_{\mathcal{C}} = \sup\{\phi(\theta) \mid \theta \in [-\tau, 0]\}$ for $\phi \in \mathcal{C}$, \mathcal{C} is a Banach space. The reader is referred to the texts [6, 10] for more background on the theory of DDEs.

Now, let G be a group. We say that system (3) is G -equivariant if there is a representation ρ of G such that for $g \in G$, we have

$$g \cdot \mathbf{F}(\mathbf{X}(t), \mathbf{X}(t - \tau)) = \mathbf{F}(g \cdot \mathbf{X}(t), g \cdot \mathbf{X}(t - \tau)) \quad (4)$$

where for ease of notation we use the symbols $g \cdot \mathbf{X}(t)$ to denote $\rho(g)\mathbf{X}(t)$. We also call G as a *symmetry group* of system (3). The equivariance condition given in Eq. (4) means that if $\mathbf{X}(t)$ is a solution to system (3), then so is $g \cdot \mathbf{X}(t)$. To see this, suppose that $\mathbf{X}(t)$ is a solution to system (3) and assume that system (3) satisfies the equivariance condition (4). Then, we have $\frac{d}{dt}[g \cdot \mathbf{X}(t)] = g \cdot \frac{d}{dt}\mathbf{X}(t) = g \cdot \mathbf{F}(\mathbf{X}(t), \mathbf{X}(t - \tau)) = \mathbf{F}(g \cdot \mathbf{X}(t), g \cdot \mathbf{X}(t - \tau))$ and this proves our claim.

For system (2), we will encounter the following $N \times N$ matrix

$$\mathbf{L} = \begin{bmatrix} A & B & \cdots & B \\ B & A & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & A \end{bmatrix}, \quad (5)$$

where A and B are scalars. We follow a similar method from [2–4] in examining the eigenvalues of matrix \mathbf{L} . For purposes of completeness, we discuss the technique that applies to matrix \mathbf{L} . Let $\zeta = e^{2\pi i/N}$ and define, for $k = 0, 1, 2, \dots, N - 1$, the subspaces $\mathbf{V}_k = \{[v, \zeta^k v, \zeta^{2k} v, \dots, \zeta^{(N-1)k} v]', v \in \mathbb{R}\}$. Notice that the action of \mathbf{L} on \mathbf{V}_k , for $k = 0, 1, 2, \dots, N - 1$, is given by $\mathbf{L}\mathbf{v}_k = \left(A + B \sum_{j=1}^{N-1} \zeta^{jk}\right) \mathbf{v}_k$ where $\mathbf{v}_k \in \mathbf{V}_k$. This means that, for $k = 0, 1, 2, \dots, N - 1$, the eigenvalue of \mathbf{L} restricted to \mathbf{V}_k is $A + \left(\sum_{j=1}^{N-1} \zeta^{jk}\right) B$. Observe that $A + \left(\sum_{j=1}^{N-1} \zeta^{jk}\right) B = A + (N - 1)B$ when $k = 0$. Similarly, we can show that $A + \left(\sum_{j=1}^{N-1} \zeta^{jk}\right) B = A - B$ for $k = 1, 2, \dots, N - 1$. Therefore, the eigenvalue of \mathbf{L} restricted to \mathbf{V}_0 is $A + (N - 1)B$, while the eigenvalue of \mathbf{L} restricted to \mathbf{V}_k , for $k = 1, 2, \dots, N - 1$, is $A - B$.

3 Results

Our first result is to show that system (2) has symmetry group S_N . Here, S_N is the *finite symmetric group* defined over the set $\{1, 2, 3, \dots, N\}$. The elements of S_N are permutations of the first N positive integers. Let $\sigma \in S_N$ and define the action of S_N to the state variables as follows

$$\sigma \cdot [x_1(t), x_2(t), \dots, x_N(t)]' = [x_{\sigma^{-1}(1)}(t), x_{\sigma^{-1}(2)}(t), \dots, x_{\sigma^{-1}(N)}(t)]'. \quad (6)$$

Since the elements of S_N are bijections from the set $\{1, 2, 3, \dots, N\}$ to itself, we have for all $\sigma \in S_N$ the following *set equality*

$$\{1, 2, 3, \dots, N\} = \{\sigma^{-1}(1), \sigma^{-1}(2), \sigma^{-1}(3), \dots, \sigma^{-1}(N)\}. \quad (7)$$

Theorem 1 *System (2) is S_N -equivariant.*

Proof If we let $\mathbf{X}(t) = [x_1(t), x_2(t), \dots, x_N(t)]'$ and

$$\mathbf{F}(\mathbf{X}(t), \mathbf{X}(t - \tau)) = \begin{bmatrix} f_1(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ f_2(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ \vdots \\ f_N(\mathbf{X}(t), \mathbf{X}(t - \tau)) \end{bmatrix}$$

where

$$f_k(\mathbf{X}(t), \mathbf{X}(t - \tau)) = a \frac{\exp(-x_k(t - \tau))}{\sum_{n=1}^N \exp(-x_n(t - \tau))} - bx_k(t) \quad (8)$$

for $k = 1, 2, \dots, N$, then system (2) can be written as in system (3). From equations (6) and (8), we have that for any element $\sigma \in S_N$ and for $k = 1, 2, \dots, N$,

$$f_k(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) = a \frac{\exp(-x_{\sigma^{-1}(k)}(t - \tau))}{\sum_{n=1}^N \exp(-x_{\sigma^{-1}(n)}(t - \tau))} - bx_{\sigma^{-1}(k)}(t).$$

Using Eq. (7), we see that $\sum_{n=1}^N \exp(-x_{\sigma^{-1}(n)}(t - \tau)) = \sum_{n=1}^N \exp(-x_n(t - \tau))$. Hence,

$$f_k(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) = a \frac{\exp(-x_{\sigma^{-1}(k)}(t - \tau))}{\sum_{n=1}^N \exp(-x_n(t - \tau))} - bx_{\sigma^{-1}(k)}(t) \quad (9)$$

for $k = 1, 2, \dots, N$. Thus, from Eqs. (8) and (9), we obtain

$$f_k(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) = f_{\sigma^{-1}(k)}(\mathbf{X}(t), \mathbf{X}(t - \tau)) \quad (10)$$

for $k = 1, 2, \dots, N$. Consequently, for any $\sigma \in S_N$, we have

$$\begin{aligned} \sigma \cdot \begin{bmatrix} f_1(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ f_2(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ \vdots \\ f_N(\mathbf{X}(t), \mathbf{X}(t - \tau)) \end{bmatrix} &= \begin{bmatrix} f_{\sigma^{-1}(1)}(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ f_{\sigma^{-1}(2)}(\mathbf{X}(t), \mathbf{X}(t - \tau)) \\ \vdots \\ f_{\sigma^{-1}(N)}(\mathbf{X}(t), \mathbf{X}(t - \tau)) \end{bmatrix} \\ &= \begin{bmatrix} f_1(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) \\ f_2(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) \\ \vdots \\ f_N(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau)) \end{bmatrix} \end{aligned}$$

where the first equality follows from the action of S_N given in Eq. (6) while the second equality follows from Eq. (10). Therefore, for any $\sigma \in S_N$, we have $\sigma \cdot \mathbf{F}(\mathbf{X}(t), \mathbf{X}(t - \tau)) = \mathbf{F}(\sigma \cdot \mathbf{X}(t), \sigma \cdot \mathbf{X}(t - \tau))$, which is the equivariance condition given in Eq. (4). This proves that system (2) is S_N -equivariant. \square

Since system (2) has symmetry group S_N , it is natural to ask if it can have solutions with the same symmetry. Solutions $(x_1(t), x_2(t), \dots, x_N(t))$ of system (2) that are *fixed* by the symmetry group S_N satisfy, for any $\sigma \in S_N$, the condition $\sigma \cdot (x_1(t), x_2(t), \dots, x_N(t)) = (x_1(t), x_2(t), \dots, x_N(t))$. This condition forces us to have $x_1(t) = x_2(t) = \dots = x_N(t)$. If we are seeking *equilibrium* solutions of system (2) that are fixed by S_N , then we need the additional condition $\frac{d}{dt}x_n(t) = 0$ for $n = 1, 2, 3, \dots, N$. This additional condition yields $x_n(t) = a/Nb$ for $n = 1, 2, 3, \dots, N$ and the equilibrium $E^* := (\frac{a}{Nb}, \frac{a}{Nb}, \dots, \frac{a}{Nb})$. Since the parameters a, b and N are all positive, the equilibrium E^* always exists. For the rest of this paper, we shall call the equilibrium E^* as the *fully symmetric equilibrium* of system (2).

Theorem 2 *The fully symmetric equilibrium E^* of system (2), that is the equilibrium fixed by the symmetry group S_N , always exists.*

3.1 Local Stability of the Fully Symmetric Equilibrium

The local stability of the fully symmetric equilibrium E^* of system (2) can be analyzed by examining the corresponding linearized system about E^* given by $\frac{d}{dt}\mathbf{X}(t) = \mathbf{M}_0\mathbf{X}(t) + \mathbf{M}_1\mathbf{X}(t - \tau)$ where the $N \times N$ matrices \mathbf{M}_0 and \mathbf{M}_1 are as follows

$$[\mathbf{M}_0|\mathbf{M}_1] = \left[\begin{array}{cccc|cccc} -b & 0 & \cdots & 0 & -\frac{N-1}{N^2}a & \frac{1}{N^2}a & \cdots & \frac{1}{N^2}a \\ 0 & -b & \cdots & 0 & \frac{1}{N^2}a & -\frac{N-1}{N^2}a & \cdots & \frac{1}{N^2}a \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -b & \frac{1}{N^2}a & \frac{1}{N^2}a & \cdots & -\frac{N-1}{N^2}a \end{array} \right].$$

The characteristic equation corresponding to the linearized system is given by

$$\det(\Delta(\lambda)) = 0 \quad (11)$$

where $\Delta(\lambda) = \lambda \mathbf{I}_N - \mathbf{M}_0 - \mathbf{M}_1 e^{-\lambda\tau}$ and \mathbf{I}_N denotes the $N \times N$ identity matrix. If all roots of the characteristic equation (11) have negative real part, then the fully symmetric equilibrium E^* of system (2) is locally asymptotically stable (LAS) [10]. Now, let $A = \lambda + b + \frac{N-1}{N^2}ae^{-\lambda\tau}$ and $B = -\frac{1}{N^2}ae^{-\lambda\tau}$. Then, the matrix $\Delta(\lambda)$ takes the form of the matrix given in Eq. (5). This means that the problem of solving the characteristic equation (11) reduces to solving the equations $A + (N-1)B = 0$ and $A - B = 0$. In fact, we have $\det(\Delta(\lambda)) = (A + (N-1)B) \cdot (A - B)^{N-1}$, which can also be proven using mathematical induction. Moreover, the roots from equation $A + (N-1)B = 0$ are simple while those from equation $A - B = 0$ are of multiplicity $(N-1)$. Since $A + (N-1)B = \lambda + b$ and $A - B = \lambda + b + \frac{1}{N}ae^{-\lambda\tau}$, we have the following lemma.

Lemma 1 *The roots of the characteristic equation (11) are the roots of the equations $\lambda + b = 0$ and*

$$\lambda + b + \frac{1}{N}ae^{-\lambda\tau} = 0. \quad (12)$$

Lemma 1 tells us that when the time delay $\tau = 0$, the roots of the characteristic equation (11) are $\lambda = -b$ and $\lambda = -b - a/N$. Both these roots are negative since $a, b, N > 0$. Consequently, we have the following theorem.

Theorem 3 *The fully symmetric equilibrium E^* of system (2) is LAS when $\tau = 0$.*

The stability of the fully symmetric equilibrium may change once the time delay parameter τ is increased. Stability switches occur when roots of the characteristic equation (11) appear on and cross the imaginary axis [9]. That is, we first need to check if $\lambda = 0$ or $\lambda = i\omega$ ($\omega > 0$) is a root of Eq. (11). In other words, using Lemma 1, we first need to check if $\lambda = 0$ or $\lambda = i\omega$ ($\omega > 0$) is a root of the equation $\lambda + b = 0$ or of Eq. (12).

The action of the symmetry group S_N on \mathbb{R}^N decomposes \mathbb{R}^N into isotypic components $\mathbf{V}_0 \oplus \mathbf{V}_1 \oplus \cdots \oplus \mathbf{V}_{N-1}$ where the subspaces \mathbf{V}_k are the subspaces we introduced in the preliminary section. Notice that S_N acts trivially on \mathbf{V}_0 while its action on \mathbf{V}_k , for $k = 1, 2, \dots, N-1$, is non-trivial. This means that roots with zero real

part obtained from $\lambda + b = 0$ give rise to regular bifurcations, while roots with zero real part obtained from Eq. (12) give rise to symmetry-breaking bifurcations [5].

Since $\lambda = 0$ is not a root of equation $\lambda + b = 0$ nor of Eq. (12), we have the following result on the steady-state bifurcations of system (2).

Theorem 4 *Steady-state bifurcations, both regular and symmetry-breaking, will not occur in system (2).*

Furthermore, since $\lambda = i\omega$ ($\omega > 0$) is not a root of equation $\lambda + b = 0$, we obtain the following result.

Theorem 5 *Regular Hopf bifurcations will not occur in system (2).*

The non-occurrence of a symmetry-breaking steady-state bifurcation rules out asymmetric equilibrium solutions in the system, while the non-occurrence of a regular Hopf bifurcation rules out synchronous periodic solutions in the system.

3.2 Symmetry-Breaking Hopf Bifurcations

We now examine the possibility of symmetry-breaking Hopf bifurcations. Suppose now that Eq. (12) has a purely imaginary root $\lambda = i\omega_*$ with $\omega_* > 0$. Since $\lambda = i\omega_*$ satisfies Eq. (12), then $i\omega_* + b + \frac{a}{N}e^{-i\omega_*\tau} = 0$ or equivalently, we have $i\omega_* + b + \frac{a}{N}(\cos \omega_*\tau - i \sin \omega_*\tau) = 0$. Matching the real and imaginary parts yields

$$a \cos \omega_*\tau = -bN \quad \text{and} \quad a \sin \omega_*\tau = N\omega_*. \quad (13)$$

We can eliminate τ by squaring each side of the equations in (13) and then adding corresponding sides to obtain $N^2\omega_*^2 = a^2 - b^2N^2$. If $(a^2 - b^2N^2) < 0$, then Eq. (12) cannot have purely imaginary roots. In this case, the roots of the characteristic equation (11) that are in the open left-half plane when $\tau = 0$ remains in the open left-half plane even if we increase the value of the time delay τ . In other words, if $(a - bN) < 0$, then the fully symmetric equilibrium E^* of system (2) is LAS for all $\tau > 0$. On the other hand, if $(a - bN) > 0$, then $(a^2 - b^2N^2) > 0$ and we obtain

$$\omega_* := \sqrt{a^2 - b^2N^2}/N > 0, \quad (14)$$

and thus the purely imaginary roots of Eq. (12) are given by $\lambda = \pm i\omega_*$. Corresponding to these roots $\lambda = \pm i\omega_*$ of Eq. (12) is the sequence

$$\tau_n := \frac{1}{\omega_*} \left\{ \cos^{-1} \left(-\frac{bN}{a} \right) + 2\pi n \right\} = \frac{\cos^{-1}(-bN/a) + 2\pi n}{\sqrt{a^2 - b^2N^2}/N} \quad (15)$$

for $n = 0, 1, 2, \dots$, obtained from the first equation in (13) and Eq. (14). In view of the Hopf bifurcation theorem [10], we now show that the roots $\lambda = \pm i\omega_*$ of Eq. (12) that lie in the imaginary axis when $\tau = \tau_n$ move towards the open right-half plane.

Lemma 2 Let $\lambda(\tau)$ be the root of Eq. (12) satisfying $\lambda(\tau_n) = i\omega_*$, for $n = 0, 1, 2, \dots$, with τ_n given in Eq. (15). Then, $\left. \frac{d}{d\tau} \operatorname{Re}(\lambda(\tau)) \right|_{\tau=\tau_n} > 0$.

Proof Note that $\operatorname{sign} \left\{ \frac{d}{d\tau} \operatorname{Re}(\lambda(\tau)) \right\}_{\tau=\tau_n} = \operatorname{sign} \left\{ \operatorname{Re} \left(\left(\frac{d\lambda}{d\tau} \right)^{-1} \right) \right\}_{\lambda=i\omega_*}$. So we first need to compute for the quantity $(d\lambda/d\tau)^{-1}$. Differentiating with respect to τ in Eq. (12) yields $(1 + \frac{1}{N} a e^{-\lambda\tau} (-\tau)) \frac{d\lambda}{d\tau} + (\frac{1}{N} a e^{-\lambda\tau} (-\lambda)) = 0$. Hence,

$$\left(\frac{d\lambda}{d\tau} \right)^{-1} = \frac{1 - \tau(a/N)e^{-\lambda\tau}}{\lambda(a/N)e^{-\lambda\tau}} = \frac{1}{\lambda(a/N)e^{-\lambda\tau}} - \frac{\tau}{\lambda} = \frac{1}{-\lambda^2 - b\lambda} - \frac{\tau}{\lambda}.$$

since $(a/N)e^{-\lambda\tau} = -(\lambda + b)$ from Eq. (12). Thus,

$$\begin{aligned} \operatorname{sign} \left\{ \frac{d}{d\tau} \operatorname{Re}(\lambda(\tau)) \right\}_{\tau=\tau_n} &= \operatorname{sign} \left\{ \operatorname{Re} \left(\frac{1}{-\lambda^2 - b\lambda} - \frac{\tau}{\lambda} \right) \right\}_{\lambda=i\omega_*} \\ &= \operatorname{sign} \left\{ \operatorname{Re} \left(\frac{1}{\omega_*^2 - ib\omega_*} - \frac{\tau}{i\omega_*} \right) \right\} \\ &= \operatorname{sign} \left\{ \frac{\omega_*^2}{\omega_*^4 + b^2\omega_*^2} \right\}. \end{aligned}$$

Since ω_* and b are both positive, we obtain that $\left. \frac{d}{d\tau} \operatorname{Re}(\lambda(\tau)) \right|_{\tau=\tau_n} > 0$. \square

We summarize the above results in the following theorem.

Theorem 6 Let $\tau^* := \min \{ \tau_n \mid \tau_n > 0 \}$ where τ_n are given in Eq. (15).

- A. If $(a - bN) < 0$, then the fully symmetric equilibrium E^* of system (2) is locally asymptotically stable (LAS) for all time delay $\tau > 0$.
- B. If $(a - bN) > 0$, then the fully symmetric equilibrium of system (2) is LAS for all $\tau \in (0, \tau^*)$ and is unstable for $\tau > \tau^*$. At $\tau = \tau^*$, system (2) undergoes a symmetry-breaking Hopf bifurcation at the fully symmetric equilibrium E^* .

In Theorem 6A, the condition for the absolute stability of E^* is $N > a/b$. That is, if we choose the number of queues to be large enough, then each queue behaves eventually the same. In other words, in this case, the decision as to which queue to wait in is immaterial since the queue lengths will eventually become equal. Meanwhile, in Theorem 6B with $N < a/b$, the Hopf bifurcation at $\tau = \tau^*$ gives rise to periodic solutions where the queue lengths oscillate periodically and asynchronously. Hence, in this case, the decision as to which queue to wait in must be done wisely since for $\tau > \tau^*$ the queues no longer give the same experience unlike in the case where $N > a/b$.

3.3 Numerical Simulations

We now illustrate the result in Theorem 6B which provides the possibility of having asynchronous periodic solutions through a symmetry-breaking Hopf bifurcation.

Example 1 Consider system (2) with $a = 10$, $b = 1$ and $N = 4$. Here, the symmetry group is S_4 , the fully symmetric equilibrium $E^* = (2.5, 2.5, 2.5, 2.5)$ and the critical delay value $\tau^* = 0.865152$ approximately. Using the initial history $(\varphi_1(t), \varphi_2(t), \varphi_3(t), \varphi_4(t)) = (2.6, 2.7, 2.8, 2.9)$ for $t \in [-\tau, 0]$, we obtain the plots in Fig. 1 showing the switch in the stability of E^* at $\tau = \tau^*$. Further examination of the periodic solution obtained for the case when $\tau = 0.9 > \tau^*$ reveals a pattern of oscillation as shown in Fig. 2.

Recall the action of the symmetry group S_4 given in Eq. (6) with $N = 4$. Observe that the non-identity elements of S_4 do not fix the periodic solution in Fig. 2. In other words, this periodic solution has *lesser* symmetry compared to the symmetry group

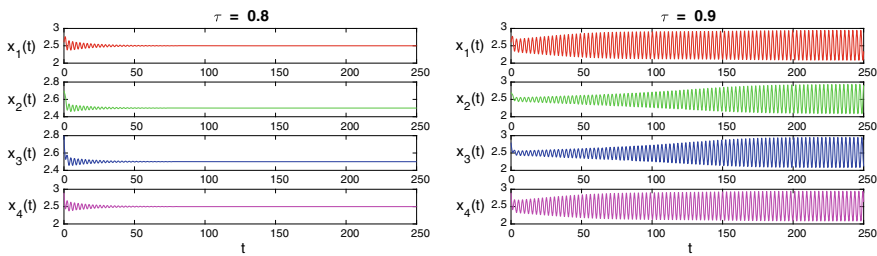
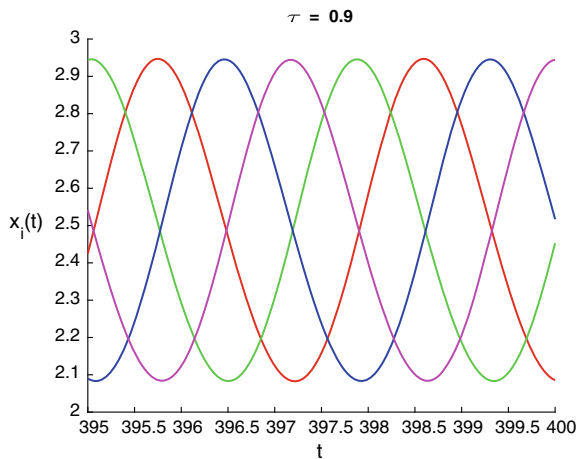


Fig. 1 A switch in the stability of the fully symmetric equilibrium E^* occurs at the critical delay value $\tau^* = 0.865152$ approximately. When $\tau = 0.8 < \tau^*$ (left panel), E^* is locally asymptotically stable while if $\tau = 0.9 > \tau^*$ (right panel), E^* is unstable

Fig. 2 Asynchronous periodic solution emerges from the symmetry-breaking Hopf bifurcation



of system (2) with $N = 4$ which is S_4 . This should not come as a surprise since this periodic solution comes from a branch or family of periodic solutions that emanates from a *symmetry-breaking* Hopf bifurcation. The only type of periodic solutions that S_4 fixed are the synchronous type. However, Theorem 5 implies the non-occurrence of such synchronous periodic solutions. Therefore, for $\tau > \tau^*$, we have the case of spontaneous symmetry-breaking since the branch or family of periodic solutions that bifurcates has smaller symmetry group than S_4 which is the symmetry group of system (2).

4 Summary and Conclusions

In this paper, we studied a generalized model describing an arbitrary number of queues. Using symmetry perspective, we showed the non-occurrence of asymmetric steady-state solutions as well as synchronous periodic solutions. This is done by ruling out symmetry-breaking steady-state bifurcations and regular Hopf bifurcations. The same technique is used to show the possibility of having asynchronous periodic solutions which is established using symmetry-breaking Hopf bifurcations. The branch or family of periodic solutions that bifurcates in this case has lesser symmetry than the symmetry group of the generalized model, which implies the occurrence of spontaneous symmetry-breaking in the system.

We showed that if the number of queues is large enough, i.e. $N > a/b$, then the fully symmetric equilibrium is asymptotically stable for all time delay τ and consequently each queue behaves eventually the same. In other words, in this case, the decision as to which queue to wait in is immaterial since the queue lengths will eventually become equal. Meanwhile, in the case where $N < a/b$, the equilibrium is only asymptotically stable when $\tau < \tau^*$. Beyond this critical value, the queue lengths oscillate periodically and asynchronously. Hence, in this case, the decision as to which queue to wait in must be done wisely since the queues no longer give the same experience unlike in the case where $N > a/b$. These additional insights on the dynamical behaviour of queues will help managers of queues to be more aware of the consequences of providing delayed queue length information to their customers.

Acknowledgements This work was funded by the UP System Enhanced Creative Work and Research Grant (ECWRG 2018-1-001). The author also acknowledges the support of the UP Baguio through RLCs during the A.Y. 2018–2019.

References

1. Brauer, F.: Absolute stability in delay equations. *J. Differ. Equ.* **69**, 185–191 (1987)
2. Buono, P.-L., Collera, J.A.: Symmetry-breaking bifurcations in rings of delay-coupled semiconductor lasers. *SIAM J. Appl. Dyn. Syst.* **14**, 1868–1898 (2015)

3. Collera, J.A.: Bifurcations of periodic solutions of functional differential equations with spatio-temporal symmetries. PhD thesis, Queen's University, Kingston (2012)
4. Collera, J.A.: Symmetry-breaking bifurcations in laser systems with all-to-all coupling. In: Bélair, J., Frigaard, I., Kunze, H., Makarov, R., Melnik, R., Spiteri, R. (eds.) *Mathematical and Computational Approaches in Advancing Modern Science and Engineering*, pp. 81–88. Springer, Cham (2016)
5. Golubitsky, M., Stewart, I., Schaeffer, D.G.: *Singularities and Groups in Bifurcation Theory*, vol. II. Springer-Verlag, New York (1988)
6. Hale, J.K., Verduyn Lunel, S.M.: *Introduction to Functional Differential Equations*. Springer, New York (1993)
7. Pender, J., Rand, R.H., Wesson, E.: Delay-differential equations applied to queueing theory. In: Stépán, G., Csernák, G. (eds.) *Proceedings of 9th European Nonlinear Dynamics Conference*, ID 62. CongressLine Ltd., Budapest (2017)
8. Pender, J., Rand, R.H., Wesson, E.: Queues with choice via delay differential equations. *Int. J. Bifurcat. Chaos.* **27**, 1730016 (2017)
9. Ruan, S., Wei, J.: On the zeros of transcendental functions with applications to stability of delay differential equations with two delays. *Dynam. Cont. Dis. Ser. A* **10**, 863–874 (2003)
10. Smith, H.: *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer, New York (2011)
11. Stewart, I.: Spontaneous symmetry-breaking in a network model for quadruped locomotion. *Int. J. Bifurcat. Chaos.* **27**, 1730049 (2017)

Algebraic Structure of the Varikon Box



Jason d'Eon and Chrystopher L. Nehaniv

Abstract The 15-Puzzle is a well studied permutation puzzle. This paper explores the group structure of a three-dimensional variant of the 15-Puzzle known as the Varikon Box, with the goal of providing a heuristic that would help a human solve it while minimizing the number of moves. First, we show by a parity argument which configurations of the puzzle are reachable. We define a generating set based on the three dimensions of movement, which generates a group that acts on the puzzle configurations, and we explore the structure of this group. Finally, we show a heuristic for solving the puzzle by writing an element of the symmetry group as a word in terms of a generating set, and we compute the shortest possible word for each puzzle configuration.

Keywords Permutation puzzles · Finite group theory · Permutation groups

1 Introduction

The 15-Puzzle is a permutation puzzle which consists of a 4×4 grid with fifteen numbered squares and one empty space that allows the pieces to slide around. Many variants of the 15-Puzzle exist, for example, by changing the size of the grid. Aside from the 15-Puzzle, one could consider the 24-Puzzle, 8-Puzzle, or the very trivial 3-Puzzle, which correspond to a 5×5 grid, a 3×3 grid, and a 2×2 grid respectively. In fact, there is no particular reason the board has to be square, so one could take the puzzle consisting of a 2×3 grid, with five movable pieces.

The focus of this paper is on a three-dimensional permutation puzzle known as the $2 \times 2 \times 2$ Varikon Box. It also consists of a $2 \times 2 \times 2$ grid with seven movable

J. d'Eon (✉)

Dalhousie University, 6299 South Street, Halifax, NS, Canada

e-mail: js348697@dal.ca

C. L. Nehaniv

University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada

e-mail: chrystopher.nehaniv@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_3

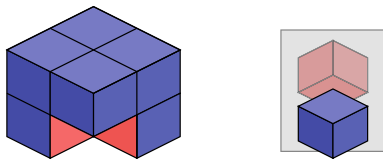


Fig. 1 The Varikon Box. The left shows an example of a solved configuration. The right shows two views of a piece inside the $2 \times 2 \times 2$ Varikon Box. It has one corner surrounded by blue faces, and the opposite corner surrounded by red faces

pieces. Each piece has three sides coloured red and three sides coloured blue, so that opposite faces are opposite colours (shown in Fig. 1). There is always one corner of the piece surrounded by red faces and one corner surrounded by blue faces, giving eight distinct orientations of a piece: one for each position of the “blue” corner (which also determines the position of the red corner). The seven pieces in the puzzle have distinct orientations. A solved state of the puzzle is a configuration where all the faces towards the outside of the puzzle are one colour, but the three faces in the core, seen through the empty space, are the opposite colour.

The outline of the paper is as follows. Section 2 is a review of the classical analysis of 15-Puzzle configurations which can be reached by valid moves. Section 3 covers which properties of the 15-Puzzle carry over to the $2 \times 2 \times 2$ Varikon Box, and describes its reachable configurations. Section 4 describes the structure of the group formed by the moves of the $2 \times 2 \times 2$ Varikon Box. Section 5 gives a heuristic for solving the Varikon Box in few moves, by writing permutations as words in terms of a generating set. Section 6 concludes with some open questions on generalizations of these puzzles.

2 Review of the 15-Puzzle

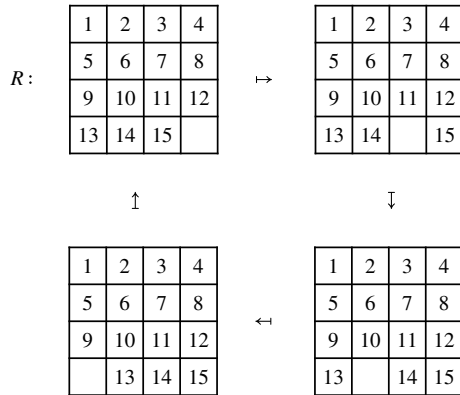
The $2 \times 2 \times 2$ Varikon Box is closely related with the 15-Puzzle, so we begin by reviewing the structure of the 15-Puzzle. In particular, we are interested in configurations which we can reach using a valid sequence of moves. Let us denote by C the set of such reachable configurations of the 15-Puzzle. For convenience, we denote the solved configuration by ι . A sequence of valid moves permutes the pieces of the puzzle, but it is not the case that every permutation of the pieces is reachable. For example, Fig. 2 shows a configuration which is well-known not to be in C .

Now take the subset C_{fix} of reachable configurations where the empty space is fixed in the bottom right corner. Configurations in C_{fix} can also be thought of as permutations in S_{15} , with respect to ι . For example, the configuration in Fig. 2 corresponds with the permutation (14, 15), since performing this permutation on the pieces of ι would yield the configuration in the figure. The following lemma was first shown in [1].

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

Fig. 2 An unsolvable configuration of the 15-Puzzle. There does not exist a sequence of moves that maps this configuration to the solved state

Fig. 3 Applying the “right” move repeatedly cycles through four 15-Puzzle configurations



Lemma 1 For every $c \in C_{fix}$, c must correspond with an even permutation.

Proof This can be seen by imagining the 4×4 grid as a black and white checkerboard. When considering moves that swap the empty space with an adjacent square, each move must change the colour that the empty space is on. If we take two configurations $c_1, c_2 \in C_{fix}$, it must take an even number of transpositions to transition from c_1 to c_2 , since the empty space begins and ends on the same colour. \square

To show that every even permutation is a reachable configuration, we adapt the proof from [2]. First, we introduce a notation for moves, which act as maps on the configurations. The definition for moves is based on the idea of sliding blocks to the right or left, as well as up or down. Unfortunately, not all moves are possible on all configurations. If the empty space is on the far left side of the grid, there is no piece to the right which can be moved to fill the space. To fix this issue, we define a “right” move, denoted R , to either mean sliding a block to the right to fill the space, or if the space is on the far left, it means to slide the entire row to the left. Figure 3 shows that under this definition, R^4 is equivalent to the identity map on C , and R^3 is what we might consider a “left” move. Similarly, we can define U to be the “up” move, which slides a piece up, effectively moving the empty space down. Using this notation, moves can be written as a sequence of R ’s and U ’s.

The diagram on the left in Fig. 4 is in C_{fix} , as it can be obtained by applying RU^3R^3U to ι , which is indicated by multiplication. Using the convention that moves are applied from left to right, this sequence corresponds with the permutation $(1, 12,$

$$\iota \cdot (RU^3R^3U) = \begin{array}{|c|c|c|c|} \hline 1 & 2 & 3 & 4 \\ \hline 5 & 6 & 7 & 8 \\ \hline 9 & 10 & 12 & 15 \\ \hline 13 & 14 & 11 & \\ \hline \end{array} \quad \iota \cdot (U^3R) = \begin{array}{|c|c|c|c|} \hline 1 & 2 & 3 & 4 \\ \hline 5 & 6 & 7 & 8 \\ \hline 9 & 10 & & 11 \\ \hline 13 & 14 & 15 & 12 \\ \hline \end{array}$$

Fig. 4 Examples of sequences being applied to the solved configuration of the 15-Puzzle. The left shows a 3-cycle in the bottom-right quadrant, and the right shows the set-up sequence, U^3R , being applied to ι

15). In order to show that every even permutation is reachable, we will use the fact that A_{15} is generated by the 3-cycles, $(11, 12, i)$ for all $i \in \{1, \dots, 15\}$ other than 11 and 12.

We start by performing a set-up sequence, U^3R , to the solved configuration, which we will undo later (Fig. 4). Following the set-up, we can then swap the empty space with the following sequence of pieces: $7 \rightarrow 8 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 1 \rightarrow 5 \rightarrow 6 \rightarrow 10 \rightarrow 9 \rightarrow 13 \rightarrow 14 \rightarrow 15 \rightarrow 7$. Repetitions of this cycle will replace 15 in the bottom-right quadrant with any other piece in this sequence, while fixing 11 and 12. Altogether, this is written:

$$\sigma_n = (U^3R)(U^3R^3U^3R^3UR^3URUR^2U^3)^n(U^3R)^{-1}, \quad (1)$$

where $n \geq 0$. Therefore, over all distinct choices of n , the sequence:

$$\sigma_n(RU^3R^3U)\sigma_n^{-1}, \quad (2)$$

will correspond with permutations of the form $(11, 12, i)$, for all i except $i = 11, 12$. By the above arguments, we have the following theorem.

Theorem 1 C_{fix} corresponds precisely with even permutations of the fifteen pieces.

A similar argument applies to any fixed position of the empty space. Therefore, the number of reachable configurations is $16 \cdot |A_{15}| = \frac{16!}{2}$. This is different than saying the reachable configurations are even permutations of the 16 squares. Rather, when the empty square is an even (or respectively, odd) number of swaps away from the bottom-right corner, then the configuration is reachable if and only if the permutation is even (respectively odd).

3 Reachable Configurations of the Varikon Box

In this section, we describe some previously known results for the $2 \times 2 \times 2$ Varikon Box [3], and provide proofs for these facts, by generalizing the 15-Puzzle. First, we note that the $2 \times 2 \times 2$ Varikon Box is precisely a three-dimensional variant of

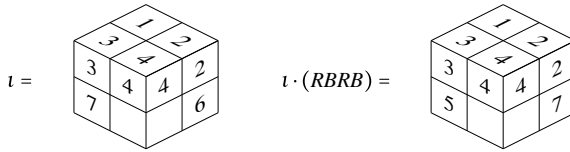


Fig. 5 Left: by labelling the pieces of the $2 \times 2 \times 2$ Varikon Box, we can define the labelling that represents the solved configuration, denoted by ι . Right: an example of a 3-cycle on the bottom half of the $2 \times 2 \times 2$ Varikon Box

the 15-Puzzle. If we choose to try solving the puzzle by making all the outer faces blue, there would be only one candidate solution, which consists of matching the blue corner of each piece with the respective corner of the puzzle. Therefore, at first glance, there would appear to be two solutions: whether we choose to put blue or red on the outer faces.

Lemma 2 *Given a fixed starting configuration, the $2 \times 2 \times 2$ Varikon Box has exactly one solution.*

Proof On each individual piece, the red corner and blue corner must be opposite from each other. Transitioning between candidate solutions would mean swapping every piece with the contents of the opposite corner, which is an even permutation. However, any sequence of moves that takes the empty space to the opposite corner will involve an odd number of swaps. Therefore, it is not possible to transition between the two candidate solutions, making only one possible to reach. \square

According to Lemma 2, we can numerically label the pieces and define a solved configuration in terms of the labelling. Let us denote the set of reachable configurations of the $2 \times 2 \times 2$ Varikon Box by V . We will reuse ι to indicate the solved configuration, shown in Fig. 5.

Lemma 1 extends to the three-dimensional case. Let V_{fix} be the set of configurations where the empty space is in its solved position. If $v \in V_{fix}$, it must correspond with an even permutation of the 7 pieces: a fact which we already used in the proof of Lemma 2. To prove that every even permutation is in V_{fix} , we show that every cycle of the form $(5, 6, i)$ is in V_{fix} , when $i \neq 5, 6$, as this will generate A_7 .

To describe sequences of swaps on the Varikon Box, we need three generators: R, U, B (right, up, and back, respectively), which act as according to Fig. 6. We define R^2, U^2 , and B^2 to be the identity map to fix the issue of certain moves being impossible given the position of the empty space. To get the permutation $(5, 6, 7)$, one can perform sequence $RBRB$ (Fig. 5), but we can also replace the 7 with any of the other pieces, by swapping the empty space with $4 \rightarrow 2 \rightarrow 1 \rightarrow 3 \rightarrow 7 \rightarrow 4$. By repeating the cycle, we can replace 7 with any other piece, in order to perform $(5, 6, i)$ for other i .

Therefore, we can extend Theorem 1 to the $2 \times 2 \times 2$ Varikon Box, since configurations in V_{fix} correspond with even permutations of the 7 pieces. By symmetry, we can conclude that there are $\frac{8!}{2} = 20,160$ reachable configurations, since for every position of the empty space, we can perform any even permutation on the 7 pieces.

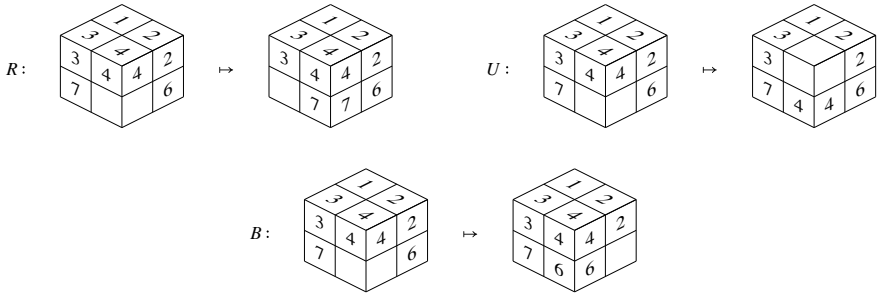


Fig. 6 The three possible directions of movement for the $2 \times 2 \times 2$ Varikon Box

4 Group Structure of the $2 \times 2 \times 2$ Varikon Box

We begin this section by observing that sequences of R , U , and B give rise to a group-like structure.

Proposition 1 *Let s_1, s_2 be two sequences of R, U , and B . We say $s_1 = s_2$ if for all $c_1, c_2 \in V$, $s_1 : c_1 \mapsto c_2$ if and only if $s_2 : c_1 \mapsto c_2$. Then with respect to composition, the set of sequences form a group and the mapping on the configurations is equivalent to a group action.*

Proof Composition is associative and concatenating two sequences will produce another valid sequence. The empty sequence satisfies the properties of the identity. Every sequence is invertible, since each element of the generating set $\{R, U, B\}$ is an involution. The group operation is well-defined, since if $x = y$ and s another sequence, then $xs = ys$, since for all $c \in V$, x and y map c to the same configuration, and performing additional moves will maintain equality. The mapping on configurations is clearly a group action, since the empty sequence leaves all configurations untouched, and the group multiplication is defined to be compatible with the action. \square

We now investigate the structure of this group, which we call G , and we show how to reduce it to a structure that will help us solve the $2 \times 2 \times 2$ Varikon Box. First, note that the stabilizer of ι is trivial and the action of the group is transitive, which implies that $|G| = 20,160$.

Interestingly, if we restrict G to the subgroup of sequences involving only R and U , we get a copy of D_6 , since $R^2 = e$, $(RU)^6 = e$, and $RU \cdot R = R \cdot (RU)^{-1}$. For any given configuration, this gives a local picture around the configuration, since by alternating any two of R, U , and B , we obtain a copy of D_6 , pictured in Fig. 7.

To help break down the size of the group, consider the group homomorphism, $\varphi : G \rightarrow (\mathbb{Z}_2)^3$, where for $g \in G$ the components of $\varphi(g)$ correspond with the counts modulo 2 of R 's, U 's, and B 's in g respectively. For example:

$$\varphi(RUBUBR) = (0, 0, 0), \tag{3}$$

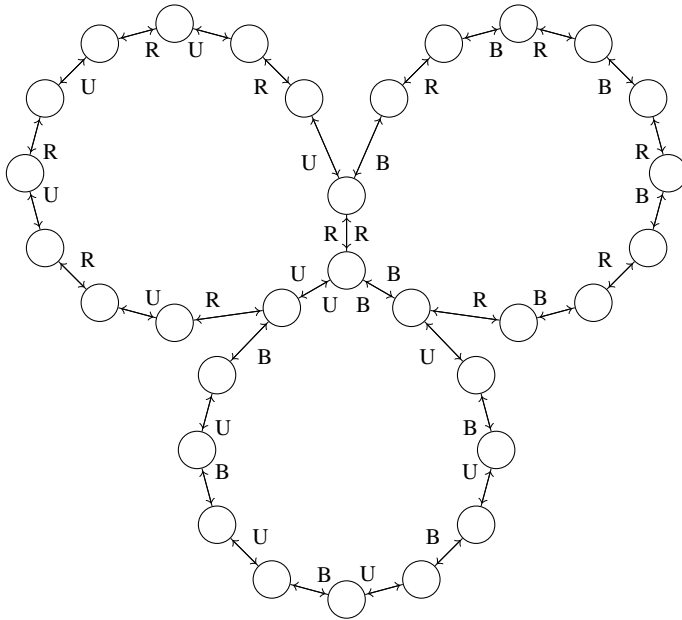


Fig. 7 Centered at a particular configuration, if one alternates between R and U , between R and B , or between U and B , we get three copies of the dihedral group of order 12

which also implies that $RUBUBR$ fixes the empty space. It is clear to see that this is a well-defined group homomorphism by properties of modular arithmetic since each letter toggles the position of the empty space in a different dimension. Consider $K = \ker \varphi$, which is a normal subgroup. By our definition of φ , K must correspond with sequences of moves which fix the empty space. By the extension of Lemma 1, $K \cong A_7$, as it acts like A_7 on the configurations in V_{fix} . Given that K is normal in G , the product, $K\langle R \rangle$, is a subgroup of G , and since $R \notin K$, we get that $|K\langle R \rangle| = 5,040$. This subgroup will be a key piece of the decomposition of G .

On the other hand, consider Z , the center of G . Computationally, we verified that $|Z| = 4$.¹ The configurations produced by applying these elements to ι are shown in Fig. 8. One can easily verify by inspection that the intersection of $K\langle R \rangle$ and Z is trivial, as no element in $K\langle R \rangle$ will move the empty space far enough to reach the non-trivial configurations in Fig. 8. Furthermore,

$$|K\langle R \rangle Z| = \frac{|K\langle R \rangle| \cdot |Z|}{|K\langle R \rangle \cap Z|} = \frac{5040 \cdot 4}{1} = 20160 = |G|. \tag{4}$$

¹The nontrivial elements can be given by the sequences: $(RU)^2(RB)^2UB(RB)^2UBRB$, $(RU)^2RB(RU)^2(BU)^2RURB$, and $(RU)^2BUBRBUBUBR(BU)^2$.

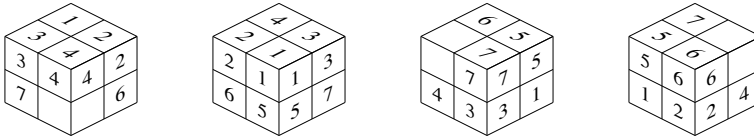


Fig. 8 The configurations obtained by applying elements of the center, Z , to ι . They correspond with ι itself, and 180° rotations of the entire box, pivoting around the U , B , and R axes

Since $K \langle R \rangle Z \leq G$, then $K \langle R \rangle Z = G$. Therefore, since $K \langle R \rangle \cap Z$ is trivial and Z commutes with $K \langle R \rangle$, we have that $G \cong K \langle R \rangle \times Z$. By determining the structure of these components, we will then obtain the full structure of G .

Lemma 3 $K \langle R \rangle \cong S_7$, where K is the kernel of the group homomorphism φ .

Proof K is normal in $K \langle R \rangle$ and $K \cap \langle R \rangle = \{e\}$, so $K \langle R \rangle = K \rtimes \langle R \rangle \cong A_7 \rtimes \mathbb{Z}_2$.² It is well known that $A_7 \rtimes \mathbb{Z}_2$ is isomorphic to either $A_7 \times \mathbb{Z}_2$ or S_7 , so it remains to show the former is false. If it held, then $K \langle R \rangle$ would contain an element kR of order 2, with $k \in K$, commuting with all of $K \langle R \rangle$. Since Z is the center, then kR commutes with all of $G = K \langle R \rangle Z$, and thus $kR \in Z$. This is a contradiction, since $K \langle R \rangle \cap Z$ is trivial. This gives us $K \langle R \rangle \cong S_7$. \square

Lemma 4 The center Z of the Varikon box group is a Klein four-group $(\mathbb{Z}_2)^2$.

Proof Note by Fig. 8 that applying a 180° rotation of the whole Varikon Box to any configuration maintains the numbers aligned along the U , B , or R axes. From this it is clear that sequences yielding these configurations commute with every other sequence, and each non-trivial element has order 2. It follows that $Z \cong (\mathbb{Z}_2)^2$. \square

Theorem 2 The group of the Varikon box G is isomorphic to $S_7 \times (\mathbb{Z}_2)^2$.

Proof This follows trivially from Lemmas 3 and 4. \square

5 The A_5 and A_6 Shortest Word Problem

One way to proceed toward a solution heuristic is by limiting the configurations to reduce the size of the problem. In practice, it is simple to locate the piece belonging in the position opposite of the empty space (the piece labelled 1) and solve it. If we only consider the configurations in V_{fix} where piece 1 is in the solved position, then we could solve the remaining pieces with $(3, 4, 7)$, $(2, 4, 6)$, and $(5, 6, 7)$, by alternating between any two of R , U , and B . With these restrictions, we can visualize

²The structure of $K \rtimes \langle R \rangle$ is given by the automorphism ϕ_R of K defined by $\phi_R(k) = RkR^{-1}$. That is, for $k_1, k_2 \in K$ and $r_1, r_2 \in \langle R \rangle$, multiplication is defined as $(k_1, r_1) \cdot (k_2, r_2) := (k_1 r_1 k_2 r_1^{-1}, r_1 r_2)$.

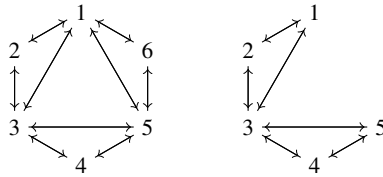


Fig. 9 A visualization of the $2 \times 2 \times 2$ Varikon Box sub-problems. The left assumes the piece opposite the empty space is in the solved position, and restricts to three 3-cycles. The right assumes two pieces are solved, and restricts to two 3-cycles

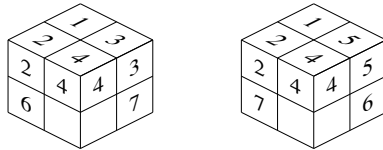


Fig. 10 Examples of the worst-cases for the two sub-problems. The left shows the A_5 sub-problem, which takes 24 moves with this method. The right shows the A_6 sub-problem, which takes 20 moves

the puzzle as just these 3-cycles on pieces 2, 3, 4, 5, 6, and 7, but for convenience, we will relabel these from 1 to 6, so that the permutations are now $(1, 2, 3)$, $(3, 4, 5)$ and $(5, 6, 1)$. Figure 9 shows a visual representation of this simplified puzzle.

One can easily verify that these 3-cycles generate A_6 . This reduces the puzzle to a word problem: given a permutation in A_6 , what is the shortest way to write it as a product of these 3-cycles and their inverses? We can solve this sub-problem by performing the inverse of this product. We can even reduce it further, by solving piece 6 of the sub-problem (which is easy in practice). This leaves us with solving the remaining 5 pieces using only the 3-cycles $(1, 2, 3)$ and $(3, 4, 5)$, which are enough to generate A_5 (a visualization of this is shown in Fig. 9).

We computed the shortest-length product of permutations in A_5 in terms of $(1, 2, 3)$ and $(3, 4, 5)$. The worst-case found was the permutation $(1, 2)(4, 5)$, with a word-length of 6. Furthermore, it was the only permutation to have this length:

$$(1, 2)(4, 5) = (3, 4, 5)^{-1}(1, 2, 3)(3, 4, 5)(1, 2, 3)^{-1}(3, 4, 5)^{-1}(1, 2, 3). \quad (5)$$

We then computed the shortest-length product of permutations in A_6 in terms of the generators $(1, 2, 3)$, $(3, 4, 5)$, and $(5, 6, 1)$. We found 46 permutations achieve the maximum word-length of 5. Which corresponds with the permutation $(2, 4, 6)$:

$$(2, 4, 6) = (1, 2, 3)(3, 4, 5)^{-1}(5, 6, 1)^{-1}(3, 4, 5)(1, 2, 3)^{-1}. \quad (6)$$

These examples are shown in their puzzle-form in Fig. 10. Setting up the A_6 sub-problem takes at most 2 moves, since the orientation of the puzzle can be freely

changed, and the worst-case scenario is when the piece labelled 1 starts adjacent to the empty space. Combining this with the worst-case for the A_6 word sub-problem, it takes at most 22 moves to solve the Varikon Box with this method, assuming one can solve the shortest word problem. This is comparable to the known worst-case which is 19, found by a brute-force method [3].

6 Conclusion and Future Work

We have analyzed the $2 \times 2 \times 2$ Varikon Box by describing moves of the puzzle as a group action on its configurations. The group associated with sequences of moves has an order equal to the number of reachable configurations, and is isomorphic to $S_7 \times (\mathbb{Z}_2)^2$. Additionally, there exist larger versions of the Varikon Box (for example, $3 \times 3 \times 3$ and $4 \times 4 \times 4$). It remains to be seen if a similar analysis can be applied to an $n \times n \times n$ Varikon Box. More abstractly, one could consider higher-dimensional variants: that is, a group generated by $\{X_1, \dots, X_k\}$, where $X_i^n = e$ for each i , which could have further applications to discrete dynamical systems in general.

Acknowledgements This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding ref. RGPIN-2019-04669, and the University of Waterloo.

References

1. Johnson, W.W., Story, W.E.: Notes on the “15” Puzzle. *Am. J. Math.* **2**(4), 397–404 (1879)
2. Mulholland, J.: *Permutation Puzzles: A Mathematical Perspective*, [self-published]. <https://www.sfu.ca/~jtmulhol/math302/notes/permutation-puzzles-book.pdf> (2019). Accessed 2 May 2020
3. Scherphuis, J.: Jaap’s Puzzle Page: Varikon Box $2 \times 2 \times 2$ / The Minus Cube. <https://www.jaapsch.net/puzzles/varikon2.htm> (2015). Accessed 2 May 2020

A Bestiary of Transformation Semigroups for the Holonomy Decomposition



Attila Egri-Nagy and Chrystopher L. Nehaniv

Abstract Here we give a selection of instructional examples of holonomy decompositions of finite transformation semigroups. These are edge cases that can be used in verifying computational implementations, as counterexamples in learning the holonomy method of the Krohn-Rhodes Theorem, and they are also sources of open problems.

Keywords Automata theory · Computer algebra · Transformation semigroup · Hierarchical decomposition · Holonomy

1 Introduction

The Krohn-Rhodes theorem decomposes a finite discrete dynamical system hierarchically into simpler components, namely as a cascade of levels of parallel flip-flops and permutation groups augmented by reset maps [9]. Computationally tractable implementations of this mathematical theorem are now available in computer algebra systems [4, 6, 7]. These rely on the holonomy decomposition method of proving the Krohn-Rhodes theorem [1, 3, 5, 8, 12], which studies the structure and covering relationships of image sets and their permutators.

The development of the computational implementation of the holonomy decomposition was not a straightforward process. This is often the case with any software project, and the open mathematical questions about the algorithm added more difficulty. The computational exploration, the testing and debugging cycles produced an interesting set of example transformation semigroups. These examples became test cases for the software package. According to the now standard continuous integration

A. Egri-Nagy (✉)
Akita International University, Akita, Japan
e-mail: egri-nagy@aiu.ac.jp

C. L. Nehaniv
University of Waterloo, Waterloo, ON, Canada
e-mail: cnehaniv@uwaterloo.ca

© Springer Nature Switzerland AG 2021
D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_4

method in software engineering, every further change in the code is tested against these examples. They include illustrative examples, and examples that cover ‘edge cases’ and exhibit unexpected features possible in discrete dynamical systems. Here, we will analyse these semigroups in order to shed light on the inner workings of the decomposition algorithm. This is useful for studying the holonomy algorithm, since the examples do not just help the software development, but they can also safeguard against the usual misunderstandings of the method. This educational perspective can also turn around the decomposition process. In a scientific scenario we have an automata model we want learn about through decomposition. But studying peculiar examples, in a way, becomes the quest for ‘engineering’ interesting decompositions, to produce decompositions with certain properties. This is a good source of open problems for further research.

Transformation Semigroups. A *transformation semigroup* (X, S) captures the concept of change in a rigorous and discrete way. It consists of a set of *states* X (analogous to *phase space*), and a set S of transformations of the state set, $s : X \rightarrow X$ acting by $x \mapsto x \cdot s$, that is closed under the associative operation of function composition. Finite state automata (without specifying initial and accepting states) and transformation semigroups are essentially the same concept, since a fixed generating set for a transformation semigroup can be considered as a set of input symbols. Writing $s_1 s_2 \in S$ for the composite function $s_1 \in S$ followed by $s_2 \in S$, we have $x \cdot (s_1 s_2) = (x \cdot s_1) \cdot s_2$, giving a (right) *action* of S on X . This extends to an action on the subsets $P \subseteq X$, by letting $P \cdot s = \{x \cdot s \mid x \in P\}$, and of special interest for us in the holonomy method are certain P where an s permutes $P \cdot s = P$ non-trivially. Transformation semigroups are general enough to model a wide range of processes. All we need is to have a strong structure theorem for them.

Decompositions. A fundamental technique of the scientific method is *decomposition*. We identify the building blocks of a system, and determine how these components work together to build the system. The simpler components are easier to understand, and we gain more understanding from the decomposition if the connections are somehow limited. When the information goes only in one direction, we talk about a *hierarchical* system. The least dependent component does not receive any information from others, while components deeper in the hierarchy are influenced by the building blocks above.

Krohn-Rhodes Theory. It is a remarkable result of finite semigroup theory [10], that we can always find a decomposition in a hierarchical form. There is a caveat though, we often end up building a bigger system through hierarchical composition. So instead of two systems being the same, we need to talk about *emulation*, which is in general a capability of one system producing the same dynamics as another one, not necessarily containing an exact copy. For semigroups, we say that S *divides* T , if S is a homomorphic image of a subsemigroup of T .

Algebraically, hierarchical connections are captured by wreath products. The *wreath product* $(X, S) \wr (Y, T)$ of transformation semigroups is the transformation semigroup $(X \times Y, W)$ where $W = \{(s, f) \mid s \in S, f \in T^X\}$, whose elements map

$X \times Y$ to itself as follows $(x, y) \cdot (s, f) = (x \cdot s, y \cdot f(x))$ for $x \in X, y \in Y$. Here T^X is the semigroup of all functions f from X to T (under pointwise multiplication). Note we have written $y \cdot f(x)$ for the element $f(x) \in T$ applied to $y \in Y$. The wreath product construction is associative on the class of transformation semigroups (up to isomorphism) and can be iterated for any number of components. Now we can state a main result of algebraic automata theory.

Theorem 1 (Krohn-Rhodes Theorem (informal statement)) *Every finite semigroup S is a divisor of a wreath product of its building block components. The groups in the components can be taken to be divisors of S itself.*

This is analogous to the Jordan-Hölder Theorem in group theory, but there we can use embedding instead of division.

2 The Holonomy Method

The holonomy decomposition is one particular method for finding the building blocks of transformation semigroups and composing them in a hierarchical structure. Beyond the ideas of emulation and hierarchy, we need two more fundamental concepts: *approximation* and *compression*.

Approximation gives less information about a system in a way that the partial description does not contradict the full description. In the holonomy decomposition, we extend the action on states to be defined on sets of states. Thus, a state is approximated by a set containing it. Then, we further extend the action to chains of increasingly smaller subsets of the state set, that successively approximate a state. The hierarchical nature of the decomposition also originates in these nested sets. The technical details of the holonomy method are for putting the extended action on chains into the form of a wreath product. For the complete algorithm see [6].

To do this we need compression, that for repeated patterns stores the pattern once and then only records its occurrences. Whenever the semigroup acts the same way on different subsets, we consider those subsets equivalent and only store the action on the equivalence class representatives (compression). These representative local actions are the building blocks of the decomposition, and they are permutation groups augmented with constant maps. They can be defined by round-trips of mappings of elements of the equivalence classes. The term ‘holonomy’ is borrowed from differential geometry: a round-trip of composed bijective maps producing permutations is analogous to moving a vector via parallel transport along a smooth closed curve yielding change of the direction of the vector.

When observing or communicating a holonomy decomposition of a transformation semigroup, the *skeleton* emerged as a most efficient tool. Though it does not provide a complete description, the skeleton depicts subsets of the state set produced by the dynamics of the semigroup and certain hierarchical relations between them. In most examples we will give the generators and the skeleton. The reader is invited

to the SgpDec package [7] in the GAP [11] computer algebra system to try out the examples and get additional information.

The Skeleton and Holonomy. The set of images of the state set under the dynamics of the semigroup, extended with the state set itself and all the singleton sets, is the mathematical object underlying the holonomy decomposition. Therefore, most of our examples are about the properties of this set.

The set $\mathcal{I}_S(X) = \{X \cdot s \mid s \in S\}$ is the *image set* of the transformation semigroup (X, S) . The *extended image set* of the state set under the action of the semigroup is $\mathcal{I}'_S(X) = \mathcal{I}_S(X) \cup \{X\} \cup \{\{x\} \mid x \in X\}$. The inclusion relation (being a subset of, \subseteq) is naturally defined on $\mathcal{I}'_S(X)$. For a given non-singleton member $P \in \mathcal{I}'_S(X)$, its maximal subsets $T \in \mathcal{I}'_S(X)$ are called the *tiles* of P , and clearly P is the union of its tiles since all singletons are in $\mathcal{I}'_S(X)$. The *subduction* relation generalizes inclusion in that we also allow the sets to be moved by S .

$$P \subseteq_S Q \iff \exists s \in S^1 \text{ such that } P \subseteq Q \cdot s \quad P, Q \in \mathcal{I}'_S(X),$$

i.e., either $P \subseteq Q$ or we can transform Q to include P under the action of S . (Here S^1 denotes S with the identity transformation adjoined in case S does not already contain it.) Therefore, subduction is a generalized inclusion, i.e. inclusion is subduction under the action of the trivial monoid. We define the \equiv_S equivalence relation on $\mathcal{I}'_S(X)$ by taking subduction in both directions: $P \equiv_S Q \iff P \subseteq_S Q$ and $Q \subseteq_S P$. The *skeleton* is the extended image set with the subduction and the corresponding equivalence relation.

The *permutator group* of P is the permutation group on P generated by restricting those $s \in S^1$ for which $P = P \cdot s$ to P , and the *holonomy group* of a nonsingleton P is the induced action of the permutator group on the tiles of P . The *height* of P is the length of longest strict subduction chain from a singleton to P . The permutator groups of equivalent sets are isomorphic as permutation groups, as are their holonomy groups. This is the source of compression in the holonomy method. The holonomy decomposition theorem [1, 3, 5, 8, 12] states that (X, S) is emulated by the wreath product of direct products of holonomy groups (augmented by constant maps), arranged hierarchically according to height, with just one holonomy group for each equivalence class in the skeleton.

3 Bestiary of Examples for Image Sets in the Skeleton

Let us examine examples with attention to properties of the skeleton's image sets.

The Antithesis of an Edge Case: The Full Transformation Semigroup. Informally speaking, a *full transformation semigroup*, consisting of all possible transformations of n states, is the easiest to decompose (Fig. 1). Because of its regularity, even an incomplete implementation of the holonomy decomposition can produce the correct decomposition. All subsets of the state set are in $\mathcal{I}_S(X)$, thus there is no missing

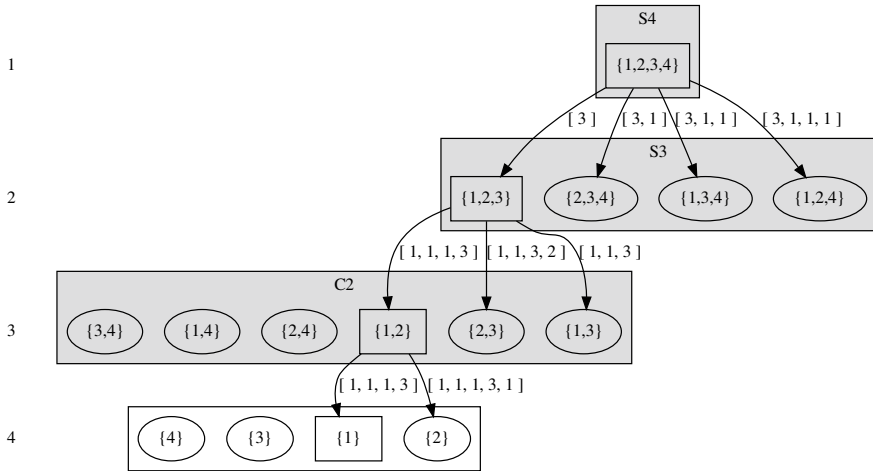


Fig. 1 The skeleton of the full transformation semigroup with 4 states. At each hierarchical level we have all the subsets of the corresponding cardinality, all in one equivalence class (depicted as rectangular boxes), acting upon by a symmetric group and constant maps. The arrows point to the tiles of a representative set. In this case they all come from the level directly below. The list in the labels is a sequence of generators that produce the tile from the representative set. The semigroup can be generated by the standard generator set: [Transformation ([2, 3, 4, 1]), Transformation ([2, 1]), Transformation ([1, 2, 3, 1])]

piece to trip over for an algorithm. The holonomy groups are symmetric and equicardinality implies subduction equivalence. A good way to understand the holonomy decomposition is to think about this base case, then see how more general examples depart from its regular structure.

All Image Sets Without Groups. The symmetric groups in a full transformation semigroup generate all subsets of the state set. For a set missing a particular single state, applying all permutations yields all the sets missing exactly one state. By iterating this process for sets missing more than one states we get all subsets of the state set.

Is it possible to generate these subsets without a non-trivial group component?

We could try to include a generator for each subset, collapsing the elements not in a given subset. However, these generators tend to combine into group components. We have to send the non-collapsed states to somewhere, and these individual maps link up to form cycles. Figure 2 shows a constructed example where we tried to add only a selected few generators, to keep the balance of generating all images but not combining into permutations. It is an open problem whether this is possible for all number states n , or just smaller cases?

No Generated Image Sets. *What are the minimal, in terms of the number of subsets, examples of holonomy decompositions?* The simplest possible dynamics is the action of a constant map (Fig. 3). It gives only a single level and just a single image set.

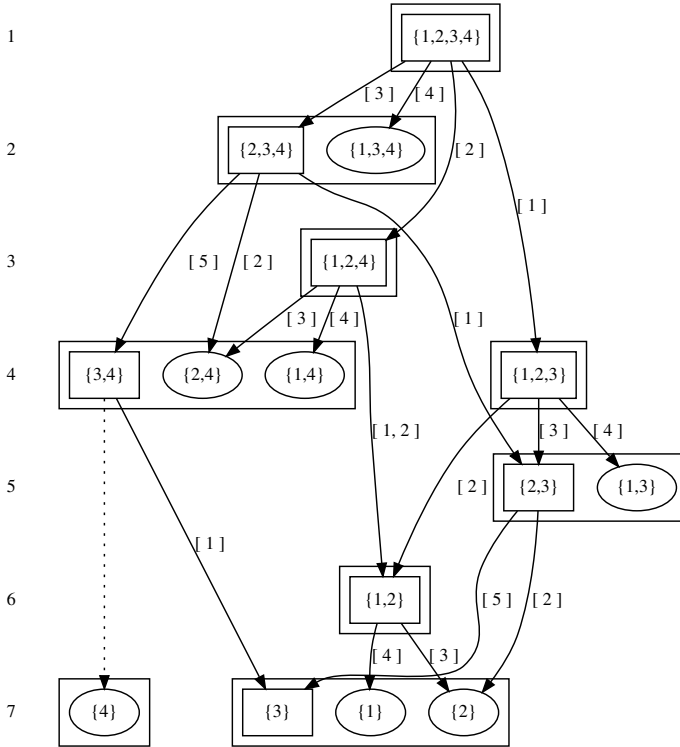


Fig. 2 A skeleton with all possible non-empty subsets as image sets but without any non-trivial group component. The generators are [Transformation ([1, 2, 3, 3]), Transformation ([1, 2, 2, 4]), Transformation ([2, 2, 3, 4]), Transformation ([1, 1, 3, 4]), Transformation ([3, 3, 3, 4])], producing image sets {1, 2, 3}, {1, 2, 4}, {2, 3, 4}, {1, 3, 4} and {3, 4}

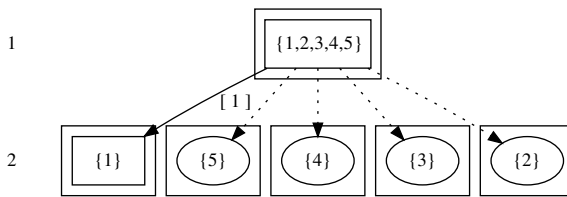


Fig. 3 Skeleton of a transformation semigroup generated by a single constant map, Transformation ([1, 1, 1, 1, 1]). The dashed arrows indicate tiles that are not images of the set being tiled

A permutation group as an input to the holonomy decomposition is also an edge case. The group action produces no proper subsets of the state set, therefore the singletons are added. These form the set of tiles as well, and the holonomy group is isomorphic to the input permutation group. This is just to say in other words that holonomy is a semigroup-, and not a group-decomposition algorithm.

Non-image Tiles. Non-image tiles are maximal subsets covering a set of states P that cannot be accessed through the dynamics of the semigroup from P itself. That is, they are tiles T of P such there is no $s \in S$ with $P \cdot s = T$. We saw these in the previous section, but there they only appeared in single-level, non-hierarchical decompositions. The real question is, *Can non-image tiles appear in non-trivial hierarchies?* The constructed example in Fig. 4 shows that this is indeed the case. This simply shows that having a subset and having an image set under the action of the semigroup are different concepts. However, understanding the holonomy decomposition often starts at the top level, where these two concept are tied together (except non-image singletons). Therefore, one might get the false impression that it is always the case.

Wide and Long Decompositions. *How wide and tall a decomposition can be for n states?* These questions were discussed in [5]. The width is easy to answer. The width can be $\binom{n}{\lfloor \frac{n}{2} \rfloor}$, and this can be realized by generating only the image sets of a middle size.

The exact limit of the number levels is an open problem. We can do a systematic analysis for transformation semigroups on 4 states, since we have all such transformation semigroups [2]. What is the length of the longest strict subduction chain?

#levels	#semigroups
1	54
2	9119
3	23953498
4	47190311
5	50321112
6	9357581
7	1238099

We have no such comprehensive data set for $n = 5$ or bigger, since the enumeration of all transformation semigroups on 5 states is beyond the current limits of computational enumeration. A better approach would be the direct study of the longest strict subduction chains.

Overtaking. The holonomy decomposition is based on nested chains of subsets. Thus, it is natural to think that as we go into deeper levels, we get smaller sets. However, there are examples of *overtaking*, a set appearing lower in the hierarchy below sets with fewer elements (Fig. 4). Since the position of a set in the skeleton is defined by strict subduction chains, this is possible. Overtaking can happen when individual tiles of a set have different length of maximal subduction chains underneath them. However, the situations in which overtakings occur have not been studied systematically.

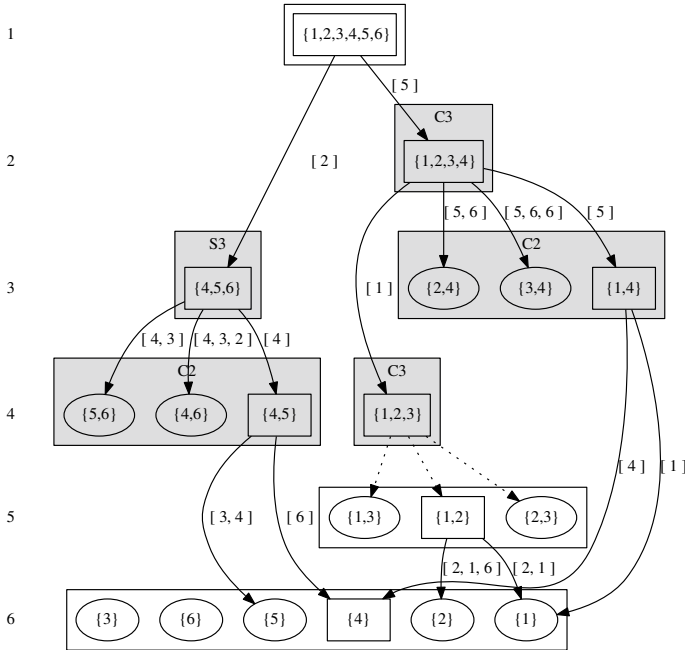


Fig. 4 The Becks transformation semigroup was constructed manually for demonstrating the existence of non-image non-singleton tiles. It also demonstrates overtaking: the set {1,4} appears above {1,2,3}. [Transformation ([1, 2, 3, 1, 1, 1]), Transformation ([4, 4, 4, 5, 4]), Transformation ([4, 4, 4, 5, 6, 4]), Transformation ([4, 4, 4, 4, 5, 5]), Transformation ([4, 4, 4, 1, 2, 3]), Transformation ([2, 3, 1, 4, 4, 4])]

4 Bestiary of Examples for Holonomy Groups

Finding the group components, ‘local pools of reversible dynamics’ is the final step of the holonomy decomposition.

Non-isomorphic Holonomy and Permutator Groups. Given a subset P in the extended image state set, we distinguish two possibly different permutation groups associated with it: The permutator group of P acting on the points of P induces the holonomy group action on the tiles of P . Therefore holonomy groups are homomorphic images of the permutator groups. For the full transformation semigroup they are isomorphic. *What is a minimal example where the permutator and the holonomy group of a subset P are not isomorphic?* The idea is to have some permutation entirely contained in a tile. Figure 5 shows how this can be constructed.

Overlapping. The holonomy decomposition can produce two separate group components on the same hierarchical level, if two sets are not equivalent. However, this does not exclude the possibility of sharing the tiles, so the sets of tiles may overlap.

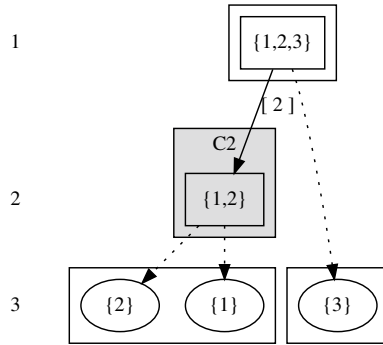


Fig. 5 Skeleton of a transformation semigroup generated by the transformations [Transformation ($[2, 1]$), Transformation ($[1, 2, 2]$)]. The top level has no non-trivial holonomy group, though the permutator group of $\{1, 2, 3\}$ is C_2 , the cyclic group of order 2. However, the top level has two tiles, and one of them contains the entire non-trivial group action

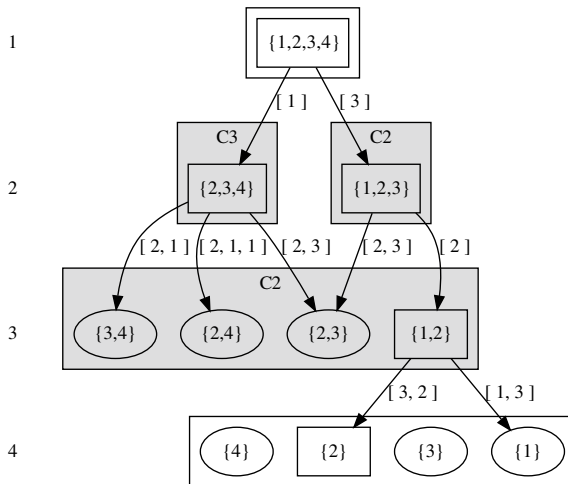


Fig. 6 Skeleton with overlapping tile set for two components on the same level. The groups C_2 and C_3 on the second level are different. Still their tile sets contain equivalent elements. The transformation semigroup is generated by [Transformation ($[4, 3, 4, 2]$), Transformation ($[1, 2, 2, 1]$), Transformation ($[3, 2, 1, 1]$)]

Figure 6 shows an example where the tiles are coming from the same equivalence class. This demonstrates, that even if a local pool of reversibility (an equivalence class) exists on a level, irreversibility can occur on the levels above.

5 Conclusion

The biggest obstacle for practical applications of computational Krohn-Rhodes methodology is the lack of comprehensive knowledge about the space of possible decompositions. The above examples could help to find the most promising ways to improve our understanding. The open problems mentioned in this paper are:

1. What is the upper bound for the height of a holonomy decomposition of a transformation semigroup on n states?
2. Is it always possible to generate all subsets of the state sets without group components?
3. What are all the ways in which the depths of subsets can be in an inverse relationship with their cardinalities.

A systematic study of the space of skeletons could shed light on these problems.

Acknowledgements This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding ref. RGPIN-2019-04669, and the University of Waterloo.

References

1. Dömösi, P., Nehaniv, C.L.: Algebraic Theory of Finite Automata Networks: An Introduction. SIAM Series on Discrete Mathematics and Applications, vol. 11. Society for Industrial and Applied Mathematics (2005)
2. East, J., Egri-Nagy, A., Mitchell, J.D.: Enumerating transformation semigroups. *Semigroup Forum* **95**(1), 109–125 (2017)
3. Egri-Nagy, A.: Algebraic hierarchical decomposition of finite state automata: a computational approach. Ph.D. thesis, University of Hertfordshire, School of Computer Science, United Kingdom (2005)
4. Egri-Nagy, A., Mitchell, J.D., Nehaniv, C.L.: Sgpdec: Cascade (de)compositions of finite transformation semigroups and permutation groups. In: *Mathematical Software ICMS 2014, LNCS*, vol. 8592, pp. 75–82. Springer (2014). https://doi.org/10.1007/978-3-662-44199-2_13
5. Egri-Nagy, A., Nehaniv, C.L.: On the Skeleton of a finite transformation semigroup. *Ann. Math. Inform.* **37**, 77–84 (2010)
6. Egri-Nagy, A., Nehaniv, C.L.: Computational Holonomy Decomposition of Transformation Semigroups (2015). [arXiv:1508.06345](https://arxiv.org/abs/1508.06345) [math.GR]
7. Egri-Nagy, A., Nehaniv, C.L., Mitchell, J.D.: SGPDEC: Software Package for Hierarchical Composition and Decomposition of Permutation Groups and Transformation Semigroups, Version 0.9.2. <https://gap-packages.github.io/sgpdec/> (2019)
8. Eilenberg, S.: *Automata, Languages and Machines*, vol. B. Academic Press (1976)
9. Krohn, K., Rhodes, J.L., Tilson, B.R.: The prime decomposition theorem of the algebraic theory of machines. In: Arbib, M.A. (ed.) *Algebraic Theory of Machines, Languages, and Semigroups*, pp. 81–125. Academic Press (1968)
10. Rhodes, J., Steinberg, B.: *The q-theory of Finite Semigroups*. Springer (2008)
11. The GAP Group: GAP—Groups, Algorithms, and Programming, Version 4.10.2. <https://www.gap-system.org> (2019)
12. Zeiger, H.P.: Cascade synthesis of finite state machines. *Inf. Control* **10**(4), 419–433 (1967). Erratum: **11**(4): 471 (1967)

Spatial Iterated Prisoner's Dilemma as a Transformation Semigroup



Isaiah Farahbakhsh and Chrystopher L. Nehaniv

Abstract The prisoner's dilemma (PD) is a game-theoretic model studied in a wide array of fields to understand the emergence of cooperation between rational self-interested agents. In this work, we formulate a spatial iterated PD as a discrete-event dynamical system where agents play the game in each time-step and analyse it algebraically using Krohn-Rhodes algebraic automata theory using a computational implementation of the holonomy decomposition of transformation semigroups. In each iteration all players adopt the most profitable strategy in their immediate neighbourhood. Perturbations resetting the strategy of a given player provide additional generating events for the dynamics. Our initial study shows that the algebraic structure, including how natural subsystems comprising permutation groups acting on the spatial distributions of strategies, arise in certain parameter regimes for the pay-off matrix, and are absent for other parameter regimes. Differences in the number of group levels in the holonomy decomposition (an upper bound for Krohn-Rhodes complexity) are revealed as more pools of reversibility appear when the temptation to defect is at an intermediate level. Algebraic structure uncovered by this analysis can be interpreted to shed light on the dynamics of the spatial iterated PD.

Keywords Dynamical systems · Group theory · Game theory · Cellular automata

I. Farahbakhsh (✉)

University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
e-mail: infarabh@uwaterloo.ca

C. L. Nehaniv

Algebraic Intelligence & Computation Laboratory, University of Waterloo,
200 University Ave W, Waterloo, ON N2L 3G1, Canada
e-mail: chrystopher.nehaniv@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_5

1 Introduction

Krohn-Rhodes (KR) theory offers powerful tools for understanding discrete-event dynamical systems (e.g. [8]). This theory decomposes any system whose dynamics can be represented as a transformation semigroup into a cascade of permutation-group layers and identity-reset (flip-flop) layers using the wreath product [5]. This yields a “coarse-to-fine graining” of both the system’s state and its dynamical transformations. The decomposition process can uncover subsystems represented by permutation groups which we call *pools of reversibility* or *natural subsystems* (see below). Algebraic structure uncovered by this analysis can be interpreted to shed light on the dynamics and complexity of many broad classes of discrete-event dynamical systems, including models found in the field of game theory.

The Prisoner’s Dilemma (PD) is an extensively studied game which explores the problem of individual versus collective profit in a simple 2-strategy model. The model is usually presented describing a situation where two partners-in-crime are imprisoned and unable to communicate. The prosecutors lack evidence and can only convict each prisoner for a lesser charge, so they offer the prisoners a deal. This deal comes as a dilemma to the prisoners as they need to choose between remaining silent or betraying their partner which would grant them a sentence lighter than that of the lesser charge, only if their partner remains silent. These two options can be represented as strategies in a game where remaining silent is referred to as cooperation and betraying the partner-in-crime is referred to as defection. This game can be applied to any situation in which there is a temptation for individuals to defect, however the net benefit of all parties is maximized if all individuals cooperate. It has been used to study the emergence of cooperation in a wide array of models in the fields of ecology, economics and psychology [1, 12, 13].

In the PD, the payoff matrix for a given player is usually represented by:

$$\begin{array}{c}
 \text{Player 2} \\
 D \quad C \\
 \text{Player 1 } D \quad \left(\begin{array}{c|c} (a, a) & (b, c) \\ \hline (c, b) & (d, d) \end{array} \right) \\
 C
 \end{array} \quad (1)$$

where every cell corresponds to each player choosing a strategy of either defect (D) or cooperate (C). The first and second elements of the tuple within each cell represent the payoff of players 1 and 2 respectively. To represent the dilemma, the payoffs are formulated with $b > d > a > c$ and to have the net payoff maximized for two cooperators, the system is further restricted such that $2d > b + c$. A common payoff matrix satisfying these conditions is:

$$\begin{array}{c}
 \text{Player 2} \\
 D \quad C \\
 \text{Player 1 } D \quad \left(\begin{array}{c|c} (1, 1) & (b, 0) \\ \hline (0, b) & (3, 3) \end{array} \right) \\
 C
 \end{array} \quad (2)$$

where $b > 3$ is a parameter referred to as the temptation to defect. When simulated as a two-player game, the players’ strategies will always converge to defection since it is the Nash equilibrium [11]. However when iterated on a spatial structure with local interactions, more complex behaviour arises, including the persistence of cooperation due to the spatial clustering of alike strategies [9, 10].

2 Spatial Algebraic Model

For the model presented in this paper, the PD is iterated on rectangular lattice with periodic boundary conditions where each cell represents a player with one of two strategies; defection represented by ‘0’ and cooperation represented by ‘1’. A small 2×3 lattice is used here due to current resource constraints of the computational algebraic analysis, but is illustrative of the general phenomena that arise.

The state space,

$$X_{bin} = \{000000, 000001, 000010, 000011, \dots, 111110, 111111\} \tag{3}$$

is made of 64 6-bit binary strings, where the i th bit from the left represents the strategy of cell i (Fig. 1). For a more notationally compact representation, this state set can also be written in decimal form with each state being the decimal integer equivalent of the binary string,

$$X = \{0, 1, 2, 3, \dots, 62, 63\}. \tag{4}$$

During each synchronous playing of the game, t (which we call a time step), each cell plays the PD with their von Neumann neighbours and gains a net payoff over all games using the payoff matrix (2). Note that since the system is a 2×3 lattice, each cell has 3 von Neumann neighbours to avoid double-counting existing neighbours with the periodic boundaries. After playing against each other, each cell updates its strategy to match that of their neighbour with maximal payoff, only if the maximal payoff is greater than their own. If two neighbouring cells with different strategies have the same maximal payoff, then cooperation is chosen.

To allow for more complexity, the model was formulated such that certain cells can have their strategy perturbed outside of t -dependent strategy evolution. We call these cells “open”. If cell i is open, there are two locally constant mappings associated with

Fig. 1 The spatial arrangement and enumeration of the cells on the 2×3 spatial Prisoner’s Dilemma lattice

1	3	5
2	4	6

that cell; d_i and c_i . These correspond to mapping the strategy of cell i to defection or cooperation respectively, regardless of the change in payoff. These mappings on the set of open cells (denoted O) make a set of locally constant mappings, resetting cell i 's strategy to either d or c but leaving others' unchanged.

$$T'_O = \{d_i, c_i\}_{i \in O}. \quad (5)$$

The set of generators for the semigroup transformations is then given by

$$T_O = T'_O \bigcup \{t\}. \quad (6)$$

Words made from elements of T_O define mappings on the set of states by applying each transformation in order from left to right. The set of transformations generated from T_O comprise a semigroup denoted by

$$S_O = \langle T_O \rangle, \quad (7)$$

and S_O acting on X gives us the transformation semigroup (X, S_O) . As T'_O is a set of locally constant mappings, it does not depend on the parameter b , however t does and its b -dependence was explored using a *python* script which also generated the semigroup mappings. In this analysis the strict inequality $b > 3 = d$ was relaxed so that $b \geq 3$. Note that for $b = 3$, the system still favours defection since although mutual cooperation has become a weak Nash equilibrium, mutual defection is still the only strict Nash equilibrium, meaning no player can change their strategy without suffering a loss in payoff. The mappings generated by the *python* script were then read into *GAP* [7] and the transformation semigroup was analyzed using the *SgpDec* package [2] to carry out a holonomy decomposition [3, 4, 8]. This yields a KR decomposition of the spatial PD model's dynamics (X, S_O) by identifying *natural subsystems*, i.e., nontrivial permutation groups whose state set is an image $X \cdot s$ of the state set X under some semigroup element $s \in S_O$ and whose permutations are the restrictions of those members of S_O which permute this set. Such an image set can be covered by the union of smaller image sets and singletons, which in turn must also be permuted by these transformations. The permutation group induced on the maximal covering sets of a natural subsystem by these sets is a *holonomy group*. See [3, 4, 8] for details. In the next section, we will be referencing *subduction*, a generalized inclusion relation defined on the collection of images together with X and the singletons. For subsets $P, Q \subseteq X$, we say P *subducts* Q if $P \subseteq Q \cdot s$ for some $s \in S$ or s the identity mapping. Mutual subduction implies isomorphism of holonomy groups, so equivalent locally reversible dynamics in the hierarchical decomposition can be compressed [3], giving insight into complexity of a dynamical system (X, S) . In the analysis and diagrams below, subduction corresponds to subset inclusion.

3 Complexity Regimes

The investigation of the iterated PD's b -dependence revealed four different regimes characterized by unique sets of transformations by t (Table 1). The complexity of each regime was explored using the Krohn-Rhodes (KR) definition of semigroup complexity [6]: KR complexity is formulated such that the complexity of a transformation semigroup (X, S) is equal to the smallest number of non-trivial groups needed for a wreath product decomposition of (X, S) . Therefore an upper bound for the KR complexity is the number of levels with non-trivial groups in the holonomy decomposition. For the remainder of this paper, upper bounds will be used when referring to KR complexity.

3.1 Regime A

Beginning with regime A ($b > 4.0$), the system has a temptation to defect so large, that t^2 acting on any state containing at least one defector will bring that state to '000000' (state 0), which we will call pure defection. (As t maps the pure cooperation state '111111' (state 63) to itself and no other states map to 63 by words generated by t , this state is left out of the subduction chains shown in Figs. 2 and 3.) The defection attractor dynamics can be visualized from subduction chain for $(X \setminus \{63\}, \langle t \rangle)$ (Fig. 2). We can choose to only examine the mappings induced by words generated by t when comparing regimes since the semigroup generated by T'_O is unchanged by the parameter b . As t^2 only maps to pure defection and the rest of the mappings in S_O are locally constant maps, there are no pools of reversibility and few levels in the holonomy decomposition, yielding a relatively trivial system.

3.2 Regime B

Regime B ($b = 4.0$) can be seen as a critical point where the system changes from regime A to C. The main difference between regimes A and B is that mixed strategy equilibria under transformation t appear in regime B. These equilibria fall under two spatial configurations up to isomorphism: "3-in-a-row" and "L-shape", shown

Table 1 Four unique regimes of the iterated Prisoner's Dilemma

Regime	Parameter range
A	$b > 4.0$
B	$b = 4.0$
C	$3.0 < b < 4.0$
D	$b = 3.0$

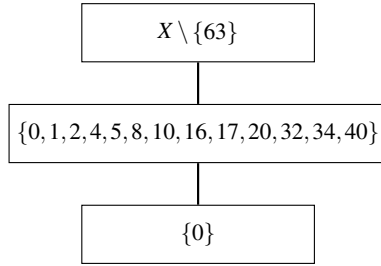


Fig. 2 Subduction chain for $(X, (t))$ with $b > 4$

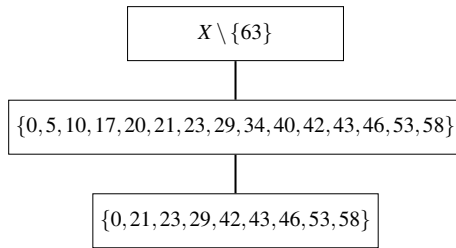


Fig. 3 Subduction chain for $(X, (t))$ with $b = 4$

1	3	5
2	4	6

1	3	5
2	4	6

(a) “3-in-a-row” strategy configuration represented by $\{21, 42\} = [21]_{\cong} \subset X$

(b) “L-shape” strategy configuration represented by $\{23, 29, 43, 46, 53, 58\} = [23]_{\cong} \subset X$

Fig. 4 Mixed strategy equilibria configuration for regimes B and C. Hatched pattern and no fill represent defector and cooperator strategies, respectively

in Fig. 4. Similar to regime A, this regime does not have non-trivial groups in the holonomy decomposition giving both regimes a KR complexity of 0.

3.3 Regime C

In regime C ($3.0 < b < 4.0$), the temptation to defect is at an intermediate level which now allows certain states to map to ones of higher cooperation with t . From the subduction chain (Fig. 5) one can see that the decreased temptation to defect removes one class of mixed strategy equilibria states, $[21]_{\cong}$. In this state, the defector’s net

Fig. 5 Subduction chain for $(X, \langle t \rangle)$ with $3 < b < 4$

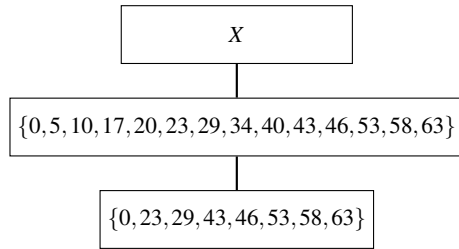


Table 2 Unique open cell configurations for 2,3 & 4 open cells, represented in grey

Open cell orientation									
KR complexity	0	0	2	0	2	4	4	4	7
Groups in holonomy decomposition	–	–	$(3, C_2)$ $(2, C_2)$	–	$(2, C_2)$	$(4, C_2)$ $(3, C_2)$ $(2, C_2)$	$(4, C_2)$ $(3, C_2)$ $(2, C_2)$	$(4, C_2)$ $(3, C_2)$ $(2, C_2)$	$(3, S_3)$ $(4, C_2)$ $(3, C_2)$ $(2, C_2)$

payoff is $b + 2$ as it receives b for playing against one adjacent cooperator as well as 2 for playing against two adjacent defectors. Since $b < 4$ in this regime, $b + 2 < 6$, the total payoff for cooperators and this state will now map to pure cooperation when acted on by t .

This regime is drastically different from regimes A and B as there are now cyclic groups in the holonomy decomposition. For these intermediate temptations to defect, the system has pools of reversibility where dynamical cycles may recur, unlike the previous regimes where any non-trivial mappings induced by words in S_O would bring the system to a state in which the previous state is inaccessible by the same transformation. This reversibility is entirely t -dependent since any words made of locally constant maps which act non-trivially on a state are by definition irreversible. Only when the temptation to defect is low enough such that an action by t can bring the system to a new state of equivalent or higher cooperation will mappings induced by words from S_O form non-trivial groups.

Additionally, the amount and distribution of open cells now play a significant role in the system complexity. In general, KR complexity increases with the number of open cells; yet for a given number of open cells, their distribution plays a significant role (Table 2). Note the open cell configuration corresponding to $T'_{123} = \{d_1, d_2, d_3, c_1, c_2, c_3\}$ has its upper bound of KR complexity reduced from 6 to 4. This is because the configuration T'_{1234} has KR complexity 4 and T'_{123} is a sub-semigroup of T'_{1234} . It follows naturally that (X, S_{1234}) emulates (X, S_{123}) and from the KR complexity axioms [6], for transformation semigroups (Y, T) and (X, S) , if (X, S) can emulate (Y, T) then the complexity of (Y, T) must be less than or equal to that of (X, S) .

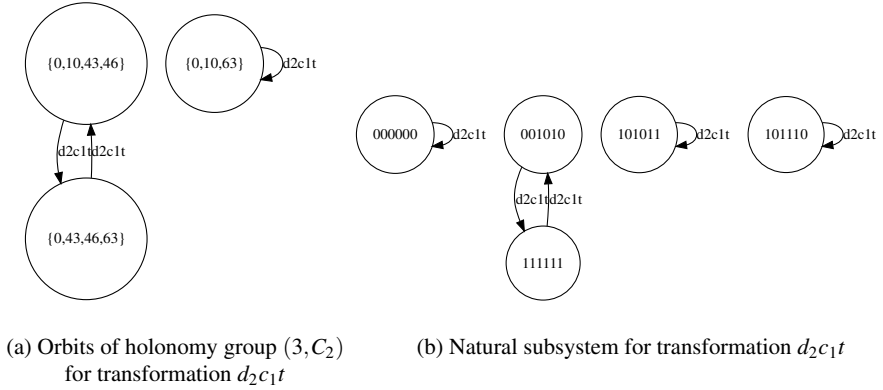


Fig. 6 Orbits and natural subsystem for a $(3, C_2)$ found in regime C with $O = \{1, 2\}$. Note states numbered 0, 10, 43, 45, and 63 in (a) correspond to strategy distributions ‘000000’, ‘001010’, ‘101011’, ‘101110’, and ‘111111’ seen in (b), respectively

Below are the orbits for a representation of the holonomy group $(3, C_2)$ found in the holonomy decomposition (Fig. 6a). This is one of two C_2 permutator groups in regime C with $O = \{1, 2\}$. The generator of this permutation group is d_2c_1t which represents mapping cell 2 to defection, cell 1 to cooperation and then letting one time step, t pass. Since the holonomy group $(3, C_2)$ shows the group action on a set of 3 subsets permuted by permutations of 5 underlying states, the exact mechanism leading to this reversibility is not immediately clear. We can gain a better understanding of the dynamics of this cyclic group by examining the orbits of the transformation d_2c_1t on specific states in these sets as shown in the natural subsystem (Fig. 6b). Most of these orbits act in an expected manner since for the two right-most orbits, the locally constant mappings do not change the state and as the states are mixed strategy equilibria for regime C , action by t does not change the state. In the left-most orbit, the behaviour is also expected since d_2c_1 effectively turns a pure defection state into one with a single cooperators which will receive the lowest payoff of its defecting neighbours, thus reverting back to defection with t .

In the orbit second from the left in Fig. 6b, the behaviour is much more interesting as the same transformation that removes four cooperators from the system, also leads it into a state of pure cooperation. For state ‘001010’, d_2c_1 acts as simply c_1 since cell 2 is already a defector. This maps the system to state ‘101010’. As shown above, this state now maps to one of pure cooperation with t . At the state of pure cooperation, d_2c_1 now acts as d_2 mapping the state to ‘101111’.

In this state, the single defector benefits from being surrounded by cooperators receiving the highest net payoff as $b > 3$ and with t , all of its neighbours switch to the defector strategy resulting in the state ‘001010’ (See Fig. 7). These pools of reversibility can offer some insight regarding the spatial configuration of strategies which lead to the persistence of cooperation. Additionally, in regime C , we see the only symmetric group $(3, S_3)$ in the holonomy decomposition (Fig. 8). Here the two

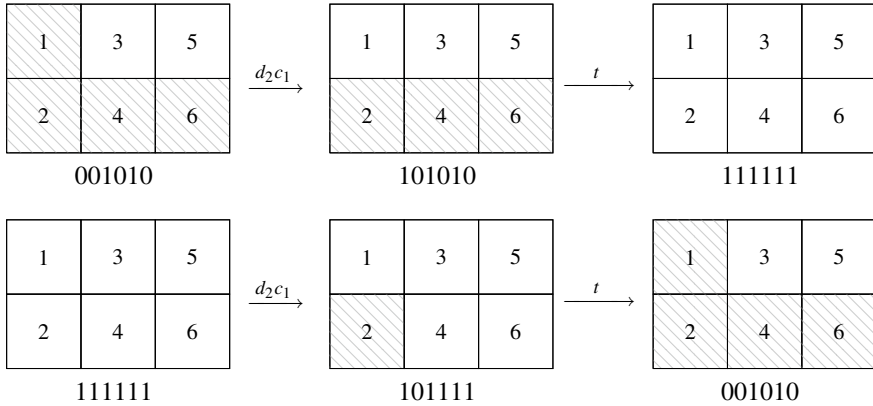


Fig. 7 Dynamics of C_2 generated by d_2c_1t

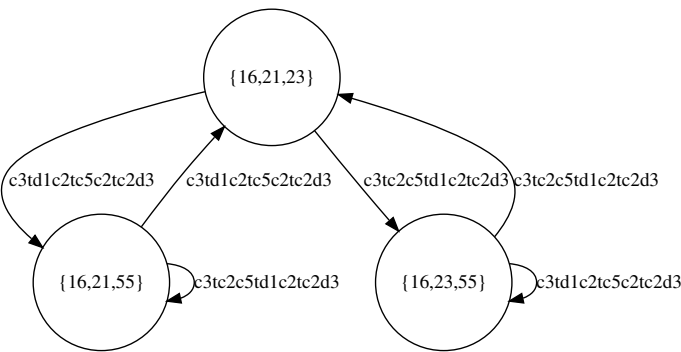


Fig. 8 Orbits of Holonomy Group $(3, S_3)$. The underlying states (strategy distributions) of the corresponding natural subsystem are $16 = '01000'$, $21 = '010101'$, $23 = '010111'$, and $55 = '101101'$

group generators are given by $c_3td_1c_2tc_5c_2tc_2d_3$ and $c_3tc_2c_5td_1c_2tc_2d_3$ and although these words are long and hard to interpret, the possibility of appearance of such non-abelian group dynamics is not an obvious result of simple iterated PD.

3.4 Regime D

In regime D ($b = 3.0$), which can be interpreted as a weak PD, the temptation to defect is very low and consequently the incentive to cooperate is highest. Due to this push towards cooperation, there are less pools of reversibility than in regime C and the highest upper bound for KR complexity is 2. From the subduction chain (Fig. 9), we see that all equilibria are mapped to by a single time step t and a new class of

Fig. 9 Subduction chain for $(X, \langle t \rangle)$ with $b = 3$

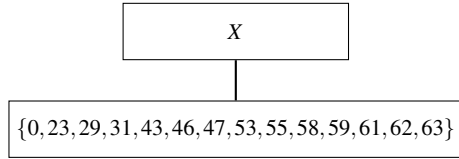


Fig. 10 New class of equilibria $\{31, 47, 55, 59, 61, 62\} = [31]_{\cong} \subset X$

1	3	5
2	4	6

equilibria emerge. (Also present are the “L-shape” equilibria with two defectors in a row we encountered above.) This new class represents a single defector, which in all previous regimes had been beneficial to the lone defector. In this regime, the temptation to defect is low enough that the system has become immune to invasion by a single defector (Fig. 10).

4 Conclusion

Representing the iterated PD as a transformation semigroup allows the holonomy decomposition to reveal qualitative differences between distinct payoff-dependent regimes. When the temptation to defect is below a threshold, the KR complexity becomes non-zero and pools of reversibility form. The number of open cells also positively influences the KR complexity, however their spatial distribution plays an equally important role. With greater computational power, it would be interesting to further explore this system with a larger number of players as well as different topologies to see how the results presented in this paper compare to larger and more complex spatial configurations. With this information, one could explore how the KR complexity varies with both spatial size and configuration, as well as with the temptation to defect. Additionally, it could lead to insights resulting in algebraic theorems for more general iterated PD systems.

All code used to generate semigroup mappings, analysis and figures is available at <https://git.io/JJcN>.

Acknowledgements This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding ref. RGPIN-2019-04669, and the University of Waterloo.

References

1. Clark, K., Sefton, M.: The sequential Prisoner's Dilemma: evidence on reciprocation. *Econ. J.* **111**(468), 51–68 (2001)
2. Egri-Nagy, A., Mitchell, J.D., Nehaniv, C.L.: SgpDec: Cascade (de)compositions of finite transformation semigroups and permutation groups. In: *International Congress on Mathematical Software*, pp. 75–82. Springer Lecture Notes in Computer Science, vol. 8592 (2014)
3. Egri-Nagy, A., Nehaniv, C.L.: Computational holonomy decomposition of transformation semigroups. [arXiv:1508.06345](https://arxiv.org/abs/1508.06345) (2015)
4. Eilenberg, S.: *Automata, Languages and Machines*, vol. B. Academic Press (1976)
5. Krohn, K., Rhodes, J.: Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Trans. Am. Math. Soc.* **116**, 450–464 (1965)
6. Krohn, K., Rhodes, J.: Complexity of finite semigroups. *Ann. Math.* 128–160 (1968)
7. Linton, S.: GAP: groups, algorithms, programming. *ACM Commun. Comput. Algeb.* **41**(3), 108–109 (2007)
8. Nehaniv, C.L., Rhodes, J., Egri-Nagy, A., Dini, P., Morris, E.R., Horváth, G., Karimi, F., Schreckling, D., Schilstra, M.J.: Symmetry structure in discrete models of biochemical systems: natural subsystems and the weak control hierarchy in a new model of computation driven by interactions. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **373**(2046), 20140,223 (2015)
9. Nowak, M.A., Bonhoeffer, S., May, R.M.: Spatial games and the maintenance of cooperation. *Proc. Natl. Acad. Sci.* **91**(11), 4877–4881 (1994)
10. Nowak, M.A., May, R.M.: Evolutionary games and spatial chaos. *Nature* **359**(6398), 826 (1992)
11. Rubinstein, A.: Finite automata play the repeated Prisoner's Dilemma. *J. Econ. Theory* **39**(1), 83–96 (1986)
12. Weitz, J.S., Eksin, C., Paarporn, K., Brown, S.P., Ratcliff, W.C.: An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proc. Natl. Acad. Sci.* **113**(47), E7518–E7525 (2016)
13. Wong, R.Y.M., Hong, Y.Y.: Dynamic influences of culture on cooperation in the Prisoner's Dilemma. *Psychol. Sci.* **16**(6), 429–434 (2005)

Oscillations and Periodic Solutions in a Two-Dimensional Differential Delay Model



Anatoli F. Ivanov and Zari A. Dzalilov

Abstract A class of two-dimensional differential systems with delay and overall negative feedback is considered. Sufficient conditions for the existence of periodic solutions are established. The instability of the unique equilibrium together with the one-sided boundedness of one of the two nonlinearities lead to the existence of periodic solutions. Systems of this type appear in various applications in engineering and natural sciences, in particular in mathematical biology and physiology as models of circadian rhythm generator and glucose-insulin regulation models in humans.

Keywords Delay differential equations · Slow oscillations · Periodic solutions · Ejective fixed point theory

1 Introduction

This paper deals with the problem of existence of slowly oscillating periodic solutions for a system of two-dimensional differential delay equations of the form

$$\begin{aligned}x'(t) &= -\alpha x(t) + f(x(t), y(t), x(t - \tau), y(t - \sigma)) \\y'(t) &= -\beta y(t) + g(x(t), y(t), x(t - \tau), y(t - \sigma))\end{aligned}\quad (1)$$

where nonlinearities f and g are continuous real-valued functions, decay rates α, β are positive, and delays τ, σ are non-negative with $\tau + \sigma := d > 0$.

Systems of type (1) appear in various applications, including physics and laser optics, physiology and mathematical biology, economics and life sciences among others. In particular, they naturally appear in physiology and mathematical biology

A. F. Ivanov (✉)
Pennsylvania State University, Dallas, PA, USA
e-mail: aivanov@psu.edu

Z. A. Dzalilov
Federation University, Ballarat, Australia
e-mail: z.dzalilov@federation.edu.au

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_6

[5, 9, 11, 14, 18, 25, 33, 35], where they serve as models of enzyme production [13, 27], of an intracellular circadian rhythm generator [30], or as glucose-insulin regulation models in humans [4, 26]. An extensive description of various applications can be found in e.g. [10, 11, 31, 35].

Sufficient conditions for the existence of periodic solutions in system (1) are established in this paper. The nonlinearities f and g are further assumed to satisfy either positive or negative feedback condition with the overall negative feedback in the system. The instability of the unique equilibrium together with a one-sided boundedness of either f or g lead to the existence of periodic solutions. The analysis of system (1) uses some of the results established for multi-dimensional systems and higher order differential delay equations [6, 20, 21].

The standard technique employed in the proof of existence of periodic solutions of autonomous equations and systems is the *Ejective Fixed Point Theory*. For details of the theory see respective chapters of monographs [8, 17]. It goes back to the classic works by Wright [34] and Jones [22–24]. Wright proved that the delay differential equation $x'(t) = -\alpha x(t-1)[1+x(t)]$ has no slowly oscillating solutions converging to zero as $t \rightarrow \infty$ when $\alpha > \frac{\pi}{2}$ (for appropriately chosen initial functions). This fact together with adapted versions of the ejective fixed point theorem by Browder [7] allowed Jones to show that the latter equation and its analogues possess slowly oscillating periodic solutions. In the years following since then the approach was further theoretically developed and formalized to the level that allowed one to prove the existence of periodic solutions to various classes of delay differential equations and systems. The basics of the ejective fixed point techniques can be found in monographs [8, 17]; some examples of application of this theory to derive periodic solutions can be given by results in papers [1, 2, 6, 15, 16, 28, 32].

In this paper we provide a general outline of establishing the existence of slowly oscillating periodic solutions for system (1); detailed mathematical exposition with complete proofs will be given in a forthcoming work. We follow the established theory of the ejective fixed point theorem as described in monographs [8, 17], with recent supplementary results obtained in [21] for systems of delay equations, as well as some further results from [6, 20]. A significant difference for the developments in this work is that our system (1) cannot be reduced to the form used in [6, 21], where the original system is transformed to the form with a single delay present in only one equation. This difference is essential and requires modified considerations and approaches. The modified elements of our approach are the selection of an appropriate cone in the phase space of system (1), deriving the relevant properties of solutions, and construction of a nonlinear map on the cone. The ejectivity is derived from establishing linear lower bounds on the functionals which are specific for this two-dimensional system. Together with the one-sided boundedness of either f or g this implies the compactness of the nonlinear map on a set of slowly oscillating initial functions, leading in turn to the existence of periodic solutions to system (1).

2 Preliminaries

2.1 Assumptions and Basics

Throughout the paper we make the following assumptions:

- (A1) (*Continuity*) Functions $f(u, v, w, z), g(u, v, w, z)$ are continuous, $f, g \in C(\mathbf{R}^4, \mathbf{R})$, the decay coefficients are positive, $\alpha > 0, \beta > 0$, and the delays τ, σ are nonnegative with $\tau + \sigma = d > 0$;
- (A2) (*Differentiability at zero*) Partial derivatives $f_u, f_v, f_w, f_z, g_u, g_v, g_w, g_z$ are continuous in a neighborhood of zero, $(u, v, w, z) \in [-\delta, \delta]^4$ for some $\delta > 0$;
- (A3) (*Overall negative feedback*) Function f satisfies the positive feedback condition

$$f(u, v, w, z) \cdot z > 0 \quad \forall (u, v, w, z) \in \mathbf{R}^4, z \neq 0, \quad (pf)$$

while function g satisfies the negative feedback condition

$$g(u, v, w, z) \cdot w < 0 \quad \forall (u, v, w, z) \in \mathbf{R}^4, w \neq 0; \quad (nf)$$

- (A4) (*One sided boundedness*) Either nonlinearity f or nonlinearity g is one-sided bounded:

$$f(u, v, w, z) \leq M > 0 \quad \text{or} \quad f(u, v, w, z) \geq -M < 0 \quad \forall (u, v, w, z) \in \mathbf{R}^4; \quad (bd)$$

(similar inequalities for g when it is one-sided bounded).

For some of the considerations of the paper a higher smoothness of the partial derivatives in (A2) may be required (see e.g. Theorem 1, (O1)). Therefore, we shall also use the following enhancement of assumption (A2):

- (A2*) Partial derivatives $f_u, f_v, f_w, f_z, g_u, g_v, g_w, g_z$ are of C^1 -class in a neighborhood of zero, $(u, v, w, z) \in [-\delta, \delta]^4$ for some $\delta > 0$.

Given initial data for each of the two components $x = \varphi(s) \in C([-\tau, 0], \mathbf{R}) := X_1, y = \psi(s) \in C([-\sigma, 0], \mathbf{R}) := X_2$ one has to solve a sequence of ordinary differential systems in order to derive the corresponding solution to system (1) for $t \geq 0$. Therefore, the natural phase space for system (1) is the direct product of two Banach spaces $X = C([-\tau, 0], \mathbf{R}) \times C([-\sigma, 0], \mathbf{R}) := X_1 \times X_2$. We shall assume that for every initial function $(\varphi, \psi) \in X$ there exists unique corresponding solution $(x(t), y(t))$ to systems (1) defined for all $t \geq 0$. Such conditions for the existence and uniqueness can be e.g. the uniform Lipschitz continuity of functions $f(u, v, w, z), g(u, v, w, z)$ in the first two variables u, v . The solution (x, y) is then constructed for $t \geq 0$ by the standard step method [3, 8, 17].

Note that due to symmetry considerations one can make the assumption in (A3) that the nonlinearity f satisfy the negative feedback condition (nf) while the nonlinearity g satisfies the positive feedback condition (pf). One can also assume, that

$\sigma \leq \tau$ is satisfied. The feedback assumptions (*pf*) and (*nf*) of (A3) imply that $(x, y) \equiv (0, 0)$ is the unique equilibrium of system (1).

System (1) includes a general cyclic system with the overall negative feedback [6, 21] as a partial case of $N = 2$, when $f(u, v, w, z) = F(z)$ and $g(u, v, w, z) = G(w)$ and functions F, G satisfy the positive and negative feedback conditions respectively in dimension one: $x \cdot F(x) > 0, x \cdot G(x) < 0, \forall x \neq 0$. An example of such systems in applications is e.g. the mathematical model for the intracellular circadian rhythm generator [30]. When $\alpha = \beta$ and $\tau = \sigma$ a partial case of system (1) was studied in papers [2, 32]. Some second order delay differential equations can also be viewed as a partial case of system (1) [1].

2.2 Linearization and Characteristic Equation

The linearized system about the equilibrium $(x, y) = (0, 0)$ is given by

$$\begin{aligned} x'(t) &= -\alpha x(t) + a_1 y(t - \sigma) \\ y'(t) &= -\beta y(t) - a_2 x(t - \tau), \end{aligned} \quad (2)$$

where $a_1 = f_z(0, 0, 0, 0) > 0$ and $a_2 = -g_w(0, 0, 0, 0) > 0$. Note that all other partial derivatives of f and g at $(0, 0, 0, 0)$ are zero due to both positive and negative feedback assumptions (*nf*) and (*pf*). The characteristic equation of the linear system (2) is found when one seeks its solutions in the exponential form $(x, y) = (x_0, y_0) \exp\{\lambda t\}$; it has the following form:

$$(\lambda + \alpha)(\lambda + \beta) + a \exp\{-d\lambda\} = 0, \quad (3)$$

where $a = a_1 a_2 > 0$ and $d = \tau + \sigma > 0$.

The transcendental equation (3) is extensively studied in several publications; we will adopt and use in this paper corresponding results from papers [1, 6].

The linear system (2) and its characteristic equation (3) determine several important properties of the original nonlinear system (1). In particular, when (3) has no real eigenvalues then all solutions to systems (1) and (2) oscillate. When (3) has a pair of complex conjugate eigenvalues with positive real part then the zero solution for both systems is unstable. See [6] for exact statements and more details.

2.3 Oscillation

We are interested in the oscillatory behavior of all solutions of system (1). Recall that a scalar continuous function $u(t)$ is said to be oscillatory on a semi-axis $[t_0, \infty)$ if there is an increasing sequence of values $t_n \rightarrow \infty$ such that $u(t_n) \cdot u(t_{n+1}) < 0$.

We shall call a solution (x, y) to the system to be oscillatory if both components $x(t)$ and $y(t)$ oscillate on the semi-axis $[t_0, \infty)$. Note that an assumption about the oscillatory behavior of one component of system (1) implies that the other component is oscillatory as well (see [6], subsection ‘‘Oscillation’’).

Sufficient conditions for the oscillation of all solutions to system (1) are given by the following statement.

Theorem 1 *Suppose that at least one of the following two conditions is satisfied:*

- (O1) *The characteristic equation (3) has no real solutions (while $(A2^*)$ holds);*
- (O2) *$ad > \max\{\alpha, \beta\}$.*

Then all solutions to system (1) oscillate about the equilibrium $(x, y) = (0, 0)$.

Part (O1) of the theorem can be derived from an analogue of Theorem 1 ([6], p. 17) when one assumes the additional smoothness properties $(A2^*)$. Part (O2) can be established along the same lines as the proof of Theorem 1 in [19] (with no principal changes). Additional oscillation criteria for delay differential equations and systems can be found in e.g. [12].

An important particular type of the oscillatory behavior is the so-called *slow oscillation*. It is associated with the size of a delay in a particular equation/system. With regard to system (1) we shall call either one of the components x or y to be slowly oscillating on a semi-axis $[t_0, \infty)$ if the distance between its any two zeros there is greater than the overall delay $d = \tau + \sigma$ in the system.

The slow oscillation is present and typical in scalar equations and systems of type (1) with the overall negative feedback. This property allows one to define an associated cone K of initial functions in the phase space X , and follow corresponding slowly oscillating solutions in forward time until they enter the cone again at some point. This return point defines a nonlinear invariant map F on cone K which fixed points correspond to periodic solutions of the original system (1). Typically the zero element $(0, 0)$ is a part of the cone, however, it produces the trivial fixed point for the nonlinear map F , as it results in constant equilibrium solution $(x, y) \equiv (0, 0)$ to system (1). One needs a second fixed point of F in order to derive a nontrivial periodic solution to the system. This is achieved by establishing the ejectivity of the trivial fixed point under the non-linear map F on the cone K .

2.4 Ejective Fixed Point Theorem

For the objectives of this paper we are adapting more general definitions and considerations of the ejective fixed point theory from [8, 17] to the partial case of two-dimensional system (1). The Banach space $X = X_1 \times X_2 = C([-\tau, 0], \mathbf{R}) \times C([-\sigma, 0], \mathbf{R})$ is the phase space of system (1). The norm $\|\cdot\|_X$ is defined as the maximum of the two norms for the Banach spaces X_1 and X_2 ; each one of the latter is defined as the supremum norm on the sets of continuous functions on initial intervals $[-\tau, 0]$ and $[-\sigma, 0]$, respectively.

Let U be a subset of X , $F : U \mapsto X$ be a mapping on U , and $x_* \in U$ be a fixed point of F . The fixed point x_* is called *ejective* if there exists its open neighborhood $G \subset X$ such that for every $x \in G \cap U$, $x \neq x_*$, there is an integer $m = m(x)$ such that $F^m(x) \notin G \cap U$.

The following statement is taken from [17] (Theorem 2.1, Sect. 11.2); its original version is given in paper [28].

Theorem 2 *Suppose \mathcal{K} is closed, bounded, convex infinite-dimensional set in X , map $\mathcal{F} : \mathcal{K} \setminus \{x_*\} \rightarrow \mathcal{F}$ is completely continuous, and x_* is an ejective fixed point of \mathcal{F} . Then there is a fixed point of \mathcal{F} in $\mathcal{K} \setminus \{x_*\}$.*

In applications of this theorem to delay differential equations the set U is usually a set of initial functions which give rise to slowly oscillating solutions (cone \mathcal{K} mentioned above). The fixed point x_* of the map \mathcal{F} is a trivial fixed point generated by a constant solution of a differential delay system. The other fixed point from $\mathcal{K} \setminus \{x_*\}$ generates a non-constant periodic solution.

3 Main Results

The following theorem is the main result of this paper

Theorem 3 *Suppose that the assumptions (A1)–(A4) are satisfied and the characteristic equation (3) has a pair of complex conjugate solutions with positive real part. Then delay differential system (1) has a nontrivial slowly oscillating periodic solution.*

The principal components of the proof are the construction of a cone $\mathcal{K} \subset X$ of initial functions, building of a nonlinear invariant map \mathcal{F} on \mathcal{K} as an appropriate shift along the corresponding solutions, showing the complete continuity of \mathcal{F} , and establishing the ejectivity of the zero fixed point of \mathcal{F} . These are outlined in the following subsections.

The ejectivity is proved under the assumption that the characteristic equation (3) has a leading pair of complex conjugate solutions $\lambda = \gamma \pm \omega i$ with the positive real part $\gamma > 0$ and the imaginary part satisfying $0 < \omega < \pi/(\tau + \sigma)$. The existence of such leading eigenvalue also implies the oscillatory behavior of all solutions of system (1).

3.1 Invariant Cone, Slow Oscillation, and Nonlinear Mapping

Consider the following cone \mathcal{K} of initial functions:

$$\mathcal{K} = \{(\varphi, \psi) \in X \mid \varphi(s) \geq 0, \varphi(s) \exp\{\alpha s\} \uparrow, s \in [-\tau, 0]; \\ \psi(s) \geq 0, \psi(s) \exp\{\beta s\} \uparrow, s \in [-\sigma, 0]\}.$$

\mathcal{K} is a closed convex set that includes the zero element $(\varphi, \psi) = (0, 0)$. The latter generates the trivial solution $(x, y) \equiv (0, 0), \forall t \geq 0$. When an initial function is not the zero element, $(\varphi, \psi) \neq (0, 0)$, then the corresponding solution $(x(t), y(t))$ is not the trivial zero solution at any $t \geq 0$.

Lemma 1 *Suppose one of the two conditions of Theorem 1 is satisfied. Assume an initial function $\mathcal{K} \ni (\varphi, \psi) \neq (0, 0)$ is given, and let $(x(t), y(t)), t \geq 0$, be the corresponding solution to system (1). Then each component $x(t)$ and $y(t)$ is slowly oscillating with the following properties holding:*

- (i) *The component $x(t)$ has an infinite sequence $\{\xi_n\}, n \in \mathbf{N}$, of simple zeros such that $\xi_{n+1} - \xi_n > d$ and $x(t) < 0 \forall t \in (\xi_{2n-1}, \xi_{2n}), x(t) > 0 \forall t \in (\xi_{2n}, \xi_{2n+1})$;*
- (ii) *The component $y(t)$ has an infinite sequence $\{\eta_n\}, n \in \mathbf{N}$, of simple zeros such that $\eta_{n+1} - \eta_n > d$ and $y(t) < 0 \forall t \in (\eta_{2n-1}, \eta_{2n}), y(t) > 0 \forall t \in (\eta_{2n}, \eta_{2n+1})$;*
- (iii) *Between any two zeros ξ_n and ξ_{n+1} of the component $x(t)$ there is exactly one zero η_{n+1} of the component $y(t)$. Likewise, between any two zeros η_n and η_{n+1} of the component $y(t)$ there is exactly one zero ξ_n of the component $x(t)$;*
- (iv) *Moreover, there is additional separation between zeros $\{\xi_n\}$ and $\{\eta_n\}$ so that the following inequalities are satisfied:*

$$\xi_n - \eta_n > \sigma \quad \text{and} \quad \eta_{n+1} - \xi_n > \tau, \quad n \in \mathbf{N}.$$

Note that Lemma 1 also holds under the assumption that the characteristic equation (3) has an eigenvalue with the positive real part. Such conditions are known explicitly (see e.g. [1] for the case $d = 1$), and can be shown to be more restrictive than those of Theorem 1.

We indicate main components of a general outline of the proof of Lemma 1.

First, each component of the solution (x, y) can be represented by an equivalent integral equation as follows. The first component $x(t)$ of the solution (x, y) of system (1) satisfies the following integral equation

$$x(t) = x_0 \exp\{-\alpha(t - t_0)\} + \int_{t_0}^t \exp\{\alpha(s - t)\} f(x(s), y(s), x(s - \tau), y(s - \sigma)) ds, \quad (4)$$

for $t \geq t_0$, where $x_0 = x(t_0)$. Likewise, the second component $y(t)$ satisfies the integral equation

$$y(t) = y_0 \exp\{-\beta(t - t_0)\} + \int_{t_0}^t \exp\{\beta(s - t)\} g(x(s), y(s), x(s - \tau), y(s - \sigma)) ds, \quad (5)$$

for $t \geq t_0$ with $y(t_0) = y_0$. Second, due to the positive and negative feedback assumptions on the nonlinearities f and g as given in assumption (A3), the increasing and decreasing nature of the expressions $x(t) \exp\{\alpha t\}$ and $y(t) \exp\{\beta t\}$ can be deduced from the representations

$$\frac{d}{dt} [x(t) \cdot \exp\{\alpha t\}] = \exp\{\alpha t\} f(x(t), y(t), x(t - \tau), y(t - \sigma)) \quad (6)$$

and

$$\frac{d}{dt} [y(t) \cdot \exp\{\beta t\}] = \exp\{\beta t\} g(x(t), y(t), x(t - \tau), y(t - \sigma)), \quad (7)$$

when $y(t - \sigma)$ and $x(t - \tau)$ are of definite sign respectively.

It follows from Eq. (5) that the component y is the first one to change the sign at some point $\eta_1 \geq 0$; moreover, $y(t)$ is decreasing on $[0, \eta_1]$ (we make a generic assumption $x(0) > 0, y(0) > 0$; other possibilities when $x(0) = 0$ or $y(0) = 0$ are analogous and eventually reduced to this one). In view of the integral equation (4) the component x remains positive on the interval $(0, \eta_1 + \sigma]$, while the component y is negative on the interval $(\eta_1, \eta_1 + \sigma]$. The component $x(t)$ is decreasing for $t \geq \eta_1$; there exists its first simple zero at $t = \xi_1 > \eta_1 + \sigma$. The component y remains negative on the interval $(\eta_1, \xi_1]$. Equations (4) and (6) next show that the component $x(t)$ is negative on $(\xi_1, \xi_1 + \sigma]$ and $x(t) \cdot \exp\{\alpha t\}$ is decreasing there. Due to the integral equation (5) the component $y(t)$ is negative in interval $[\xi_1, \xi_1 + \tau]$. In view of (7) the component y is increasing in some right neighborhood $[\xi_1 + \tau, \eta_2]$ of $\xi_1 + \tau$ where η_2 is its second simple zero. The component x remains negative on the interval $[\xi_1, \eta_2]$.

One can consider now the constructed solution (x, y) on the interval $[0, \eta_2]$ as a new initial function, an element of X . One concludes that the new initial functions $\varphi_1(s) := x(\eta_2 + s), s \in [-\tau, 0]$ and $\psi_1(s) := y(\eta_2 + s), s \in [-\sigma, 0]$ belong to the symmetric “negative set” $-\mathcal{K}$ consisting of initial functions (φ, ψ) such that $(-\varphi, -\psi) \in \mathcal{K}$. One can construct now the solution (x, y) for $t \geq 0$ in the very same way to conclude that there exists the second simple zero ξ_2 of the component x such that $\xi_2 > \eta_2 + \sigma$ and $x(t) > 0, t \in [\eta_2 + \sigma, \xi_2], y(t) > 0, t \in (\eta_2, \xi_2]$ and $y(t) \exp\{\beta t\}$ is increasing in $[\eta_2, \xi_2]$. Continuing further one shows that $x(t) > 0, t \in (\xi_2, \xi_2 + \sigma]$ and $x(t) \exp\{\alpha t\}$ is increasing there. At the same time $y(t) > 0$ and $y(t) \exp\{\beta t\}$ is increasing on the interval $(\eta_2, \xi_2 + \sigma]$.

One now defines a mapping \mathcal{F} on \mathcal{K} as follows

$$\forall (\varphi, \psi) \in \mathcal{K} : \mathcal{F}(\varphi, \psi) = (\varphi_1, \psi_1),$$

where $(\varphi, \psi) \neq (0, 0)$ and $\varphi_1(s) = x(\xi_2 + \tau + s), s \in [-\tau, 0]$ and $\psi_1(s) = y(\xi_2 + \tau + s), s \in [-\sigma, 0]$. Due to the construction described in Lemma 1 one has that $(\varphi_1, \psi_1) \in \mathcal{K}$, thus showing that \mathcal{F} maps \mathcal{K} into itself. The one-sided boundedness of either f or g implies that the derivatives of both components x and y are bounded (after the second zeros ξ_2 and η_2). Therefore the map \mathcal{F} is completely continuous, as the set $\mathcal{F}(\mathcal{K})$ of functions is bounded and uniformly continuous.

By the continuity of \mathcal{F} the zero element $(\varphi, \psi) \equiv (0, 0) \in \mathcal{K}$ is defined to be mapped into itself under \mathcal{F} , as the initial function $(\varphi, \psi) \equiv (0, 0)$ results in the zero trivial solution of system (1). Any nonzero fixed point $\mathcal{K} \ni (\varphi_*, \psi_*) \neq (0, 0)$ of map \mathcal{F} , $\mathcal{F}(\varphi_*, \psi_*) = (\varphi_*, \psi_*)$, gives rise to a nontrivial slowly oscillating periodic

solution to system (1). The ejectivity property of the trivial fixed point $(0, 0)$ guarantees the existence of such second fixed point (φ_*, ψ_*) . Its general outline is given in the next sub-section.

3.2 Ejectivity

The ejectivity of mapping \mathcal{F} is decided from certain properties of linear functionals constructed on the basis of the linearized system (2) (see [8, 17, 21] for details of general exposition about the functionals). The functionals are coming from projections on eigenspaces associated with particular eigenvalues; they are related to the Laplace transform of the linearized systems (see e.g. [1, 15, 21, 29, 34] for more details on specific cases). The functionals for system (2) turn out to be of the form:

$$L_1(x, y) = (\lambda + \beta)x(0) + (\lambda + \alpha)(\lambda + \beta) \int_{-\tau}^0 \exp\{-\lambda s\}x(s) ds \\ + a_1 \exp\{-\lambda\sigma\}y(0) + a_1(\lambda + \beta) \exp\{-\lambda\sigma\} \int_{-\sigma}^0 \exp\{-\lambda s\}y(s) ds$$

and

$$L_2(x, y) = (\lambda + \alpha)y(0) + (\lambda + \alpha)(\lambda + \beta) \int_{-\sigma}^0 \exp\{-\lambda s\}y(s) ds \\ - a_2 \exp\{-\lambda\tau\}x(0) - a_2(\lambda + \alpha) \exp\{-\lambda\tau\} \int_{-\tau}^0 \exp\{-\lambda s\}x(s) ds.$$

The two functionals are derived by using the Laplace transform $\mathcal{L}_u(\lambda) := \int_0^\infty \exp\{-\lambda t\}u(t) dt$ on the components x and y of the solutions of system (2) when the latter is subject to the Laplace transformation. Functional L_1 comes out when $\mathcal{L}_x(\lambda)$ is excluded from the algebraic system, while L_2 appears when $\mathcal{L}_y(\lambda)$ is excluded (therefore, they are equivalent). The functionals are also used to represent the projection $\Pi(\lambda)$ on the eigenspace corresponding to an eigenvalue λ (see e.g. [8, 17, 20, 21] for more details, where the relationship between the Laplace transforms of solutions and the projection Π is described and studied).

The ejectivity follows from either one of the two inequalities

$$|L_1(x, y)| \geq c_1 \|(x, y)\| \quad \text{or} \quad |L_2(x, y)| \geq c_2 \|(x, y)\|, \quad \forall (x, y) \in \mathcal{H},$$

which are in turn equivalent to the projection's boundedness away from zero

$$\sup\{|L_k(x, y)|, \|(x, y)\| = 1\} = l_k > 0, \quad k = 1, 2,$$

when $\lambda = \gamma + \omega i$ is the leading solution of the characteristic equation with the positive real part $\gamma > 0$ and the imaginary part satisfying $0 < \omega < \pi/(\tau + \sigma)$ (see e.g. [17], Theorem 2.3 (ii), p. 337). The latter is proved by using a detailed analysis of functionals L_1 and L_2 under the assumption $|(x, y)| = 1$. For example, when $|x| = 1$ then $x(0) \geq \exp\{-\alpha\tau\}$ is satisfied. One considers next the expression $L_1^* = L_1(x, y) \exp(\lambda\sigma)$ and estimates the lower bound of its imaginary part. The first term of L_1^* is estimated as $|\text{Im}\{(\lambda + \beta) \exp(\lambda\sigma)\}| \geq m_0$ for some $m_0 > 0$ independent of the particular choice of $x(s)$, $s \in [-\tau, 0]$ (when $\sigma < \tau$). The imaginary parts of the second and forth terms of L_1^* can be shown to be each positive (however, not uniformly bounded away from zero). The third term of L_1^* is pure real. Therefore, one deduces that $|L_1^*| \geq |\text{Im}(L_1^*)| \geq m_1 > 0$ is satisfied, implying also that $|L_1| \geq m_2 > 0$ is valid. The other consideration $|y| = 1$ is similar with the use of functional L_2 .

4 Conclusion

We establish sufficient conditions for the existence of non-trivial slowly oscillating periodic solutions for a new class of two-dimensional differential delay systems. Those systems are more general than some of the previously studied models such as systems with cyclic overall negative feedback [6, 21], or a model of a circadian rhythm generator [30], or a second order non-linear differential delay equation [1]; they include those mentioned as partial cases. The proof of the existence of periodic solutions follows along the lines of standard techniques of the ejective fixed point theory [8, 17]. However, the specific considerations on major steps somewhat differ from those previously employed. These distinctions require new appropriate adjustments and further developments to several points of the established theory.

Acknowledgements This work was initiated in the fall 2016 during A. Ivanov's visit and research stay at the University of Giessen, Germany, under the support of the Alexander von Humboldt Stiftung. In its final stages and during the preparation for publication A. Ivanov and Z. Dzalilov were supported by internal grants from the Federation University Australia (Ballarat, Victoria).

References

1. An der Heiden, U.: Periodic solutions of a nonlinear second order differential equation with delay. *J. Math. Anal. Appl.* **70**, 599–609 (1979)
2. Babtistini, M.Z., Táboas, P.Z.: On the existence and global bifurcation of periodic solutions to planar delay differential equation. *J. Differen. Equ.* **127**, 391–425 (1996)
3. Bellman, R., Cooke, K.L.: *Differential-Difference Equations*. Academic Press, New York/London (1963)
4. Bennett, D.L., Gourley, S.A.: Global stability in a model of the glucose-insulin interaction with time delay. *Euro. J. Appl. Math.* **15**, 203–221 (2004)

5. Berezensky, L., Braverman, E., Idels, L.: Nicholson's blowflies differential equations revisited: main results and open problems. *Appl. Math. Model.* **34**, 1405–1417 (2010)
6. Braverman, E., Hasik, K., Ivanov, A., Trofimchuk, S.: A cyclic system with delay and its characteristic equation. *Discret. Contin. Dyn. Syst. Ser. S* **13**, 1–29 (2020)
7. Browder, F.E.: A further generalization of the Schauder fixed point theorem. *Duke Math. J.* **32**, 575–578 (1965)
8. Diekmann, O., van Gils, S., Verduyn Lunel, S.M., Walther, H.-O.: *Delay Equations: Complex, Functional, and Nonlinear Analysis*. Springer-Verlag, New York (1995)
9. Duan, L., Huang, L., Chen, Y.: Global exponential stability of periodic solutions to a delay Lasota-Ważewska model with discontinuous harvesting. *Proc. AMS* **144**, 561–573 (2016)
10. Erneux, T.: *Applied Delay Differential Equations. Surveys and Tutorials in the Applied Mathematical Sciences*, vol. 3. Springer Verlag (2009)
11. Glass, L., Mackey, M.C.: *From Clocks to Chaos. The Rhythms of Life*. Princeton University Press (1988)
12. Györy, I., Ladas, G.: *Oscillation Theory of Delay Differential Equations*. Oxford Science Publications, Clarendon Press, Oxford (1991)
13. Goodwin, B.C.: Oscillatory behaviour in enzymatic control process. *Adv. Enzyme Regul.* **3**, 425–438 (1965)
14. Haderl, K.P.: *Delay Equations in Biology. Springer Lecture Notes in Mathematics*, vol. 730, pp. 139–156 (1979)
15. Haderl, K.P., Tomiuk, J.: Periodic solutions of difference differential equations. *Arch. Rat. Mech. Anal.* **65**, 87–95 (1977)
16. Hale, J.K., Ivanov, A.F.: On a high order differential delay equation. *J. Math. Anal. Appl.* **173**, 505–514 (1993)
17. Hale, J.K., Verduyn Lunel, S.M.: *Introduction to Functional Differential Equations. Applied Mathematical Sciences*. Springer-Verlag (1993)
18. Hopfield, J.: Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982)
19. Ivanov, A.F., Dzalilov, Z.A.: Oscillations in low-dimensional cyclic differential delay systems. In: Kilgour, D., Kunze, H., Makarov, R., Melnik, R., Wang, X. (eds.) *Recent Advances in Mathematical and Statistical Methods. AMMCS 2017. Springer Proceedings in Mathematics & Statistics*, vol. 259. Springer, Cham (2018)
20. Ivanov, A.F., Lani-Wayda, B.: Periodic solutions for three-dimensional non-monotone cyclic systems with time delays. *Discret. Contin. Dyn. Syst.* **11**, 667–792 (2004)
21. Ivanov, A.F., Lani-Wayda, B.: Periodic solutions for an N -dimensional cyclic feedback system with delay. *J. Differen. Equ.* **268**, 5366–5412 (2020)
22. Jones, G.: The existence of periodic solutions of $f'(x) = -\alpha f(x-1)[1+f(x)]$. *J. Math. Anal. Appl.* **5**, 435–450 (1962)
23. Jones, G.: On the nonlinear differential difference equation $f'(x) = -\alpha f(x-1)[1+f(x)]$. *J. Math. Anal. Appl.* **4**, 440–469 (1962)
24. Jones, G.: Periodic motions in Banach space and applications to functional differential equations. *Contrib. Differen. Equ.* **3**, 75–106 (1964)
25. Kuang, Y.: *Delay Differential Equations with Applications in Population Dynamics. Series: Mathematics in Science and Engineering*, vol. 191. Academic Press Inc. (2003)
26. Li, J., Kuang, Y., Mason, C.: Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two time delays. *J. Theoret. Biol.* **242**, 722–735 (2006)
27. Mahaffy, J.: Periodic solutions of certain protein synthesis models. *J. Math. Anal. Appl.* **74**, 72–105 (1980)
28. Nussbaum, R.D.: Periodic solutions of some nonlinear autonomous functional differential equations. *Ann. Mat. Pura Appl.* **101**, 263–306 (1974)
29. Nussbaum, R.D.: Periodic solutions of nonlinear autonomous functional differential equations. In: *Functional Differential Equations and Approximation of Fixed Points. Summer School and Conference, University of Bonn, Bonn, 1978*, pp. 283–325. *Lecture Notes in Mathematics*, vol. 730. Springer, Berlin (1979)

30. Scheper, T., Klinkenberg, D., Pennartz, C., van Pelt, J.: A mathematical model for the intracellular circadian rhythm generator. *J. Neurosci.* **19**(1), 40–47 (1999)
31. Smith, H.: *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Series: Texts in Applied Mathematics, vol. 57. Springer-Verlag (2011)
32. Táboas, P.: Periodic solutions of a planar delay equation. *Proc Roy. Soc. Edinburgh* **116A**, 85–101 (1990)
33. Tan, Y., Zhang, M.: Global exponential stability of periodic solutions in a nonsmooth model of hematopoiesis with time-varying delays. *Math. Meth. Appl. Sci.* 2017, 10 pp. <https://doi.org/10.1002/mma.4448>
34. Wright, E.M.: A non-linear differential-difference equation. *J. Reine Angew. Math.* **194**, 66–87 (1955)
35. Wu, J.: *Introduction to Neural Dynamics and Signal Transmission Delay*. *Nonlinear Analysis: Real World Applications*, vol. 6. Walter de Gruyter & Co., Berlin (2001)

Exploring Tetris as a Transformation Semigroup



Peter C. Jentsch and Chrystopher L. Nehaniv

Abstract Tetris is a popular puzzle video game, invented in 1984. We formulate two versions of the game as a transformation semigroup and use this formulation to view the game through the lens of Krohn-Rhodes theory. In a variation of the game upon which it restarts if the player loses, we find permutation group structures, including the symmetric group S_5 which contains a non-abelian simple group as a subgroup. This implies, at least in a simple case, that iterated Tetris is finitarily computationally universal.

Keywords Ames · Krohn-Rhodes algebraic automata theory · Semigroups · Computer algebra · Holonomy

1 Introduction

Tetris is an arcade puzzle game created by Alexey Pajitnov in 1984, that has since become a worldwide cultural phenomenon [9]. It is the best selling paid-downloaded mobile game of all time, with over 100 million copies sold for cellphones [11]. It is also the most ported video game ever, according to the Guinness Book of World Records, with an estimated 65 platforms [11]. Tetris is fundamentally a polyomino stacking game. The playing field consists of a 10×20 grid, and the player is given a sequence of tetriminoes (Fig. 1), which are sets of four connected grid cells, to drop from the top of the playing field. The player can translate or rotate the shapes as they fall. If a row is filled, the row disappears, and all the blocks above that row are moved down by one row. If the blocks are stacked outside the grid, then the game

P. C. Jentsch (✉)

University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada

e-mail: pjentsch@uwaterloo.ca

C. L. Nehaniv

Waterloo Algebraic Intelligence & Computation Laboratory, University of Waterloo,

200 University Ave W, Waterloo, ON N2L 3G1, Canada

e-mail: cnehaniv@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_7

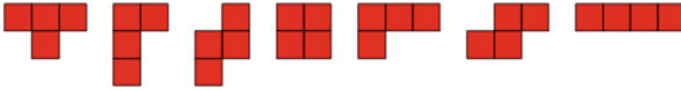


Fig. 1 Pieces available in standard Tetris, the one-sided tetrominoes [16]

is over. The object of the game is to survive as long as possible. While generally, the player is only allowed to see one or two pieces ahead, most authors consider the version where the player has access to the full sequence of pieces ahead of time. This variation is also called *offline* Tetris [4], and unless otherwise stated will be the one discussed here.

There has been considerable research into the mathematics behind Tetris, for an arcade game. Demaine et al. show that the complexity of solving many aspects of the game, such as maximizing the number of rows cleared, or the number of moves before the game ends, are NP-complete. Furthermore, they show that finding algorithms which approximate solutions to these is quite difficult [4]. Other authors have characterized optimal strategies for small subsets of the pieces [2], and characterized sequences of pieces that always cause a loss [3]. Hooeboom and Kosters [10] show that nearly any reasonable configuration of blocks is possible to construct under the Tetris rules with a suitable sequence of tetrominoes. It is also possible to represent Tetris, and other tiling games, as regular grammars, which has allowed for some enumeration of possible Tetris games [1].

We formulate the game of Tetris as a transformation semigroup, where the elements of the semigroup are transformations on the set of possible game states. Krohn-Rhodes theory [12, 13] and the related holonomy decomposition (Theorem 3) [6, 7] provide a way to decompose transformation semigroups into wreath products of finite simple groups and the flip-flop monoid (see Appendix for concepts and theorems employed here related to holonomy). Our analysis is primarily computational, and we use a package for the computer algebra system GAP called “SgpDec” [5].

2 Tetris as a Transformation Semigroup

Let P be a set of pieces, where a “piece” is a set of connected cells such as a tetromino. Let S be a semigroup generated by basic events $\sigma = (p, \xi) \in S$ consisting of a set of connected cells p , and a position $1 \leq \xi \leq n$ (although the precise limits on the position ξ depend on the width of p).

A configuration, or state, x is an element of the set of $n \times k$ board of cells, where some of the cells are filled by other pieces. An element $\sigma \in S$ acts on a configuration x by “dropping” the piece p with the *leftmost* block in the column ξ , and then applying the row removal rules in the familiar way. That is, if there is a full row of width n , it is removed, and the blocks above the row are dropped down by one. (If more full rows arise, removal and dropping is iterated.) We denote the empty game state before

any pieces have dropped by e , and denote by E the “game over” state. If the total number of cells in the stack exceeds k , then $x \cdot \sigma = E$. Furthermore, $E \cdot \sigma = E$ for all $\sigma \in S$. For $\sigma_1, \sigma_2 \in S$, define their product $\sigma_1\sigma_2$ as the transformation resulting from applying σ_1 then σ_2 in the above way.

We consider a state x in the set of possible permutations of a $n \times k$ board of cells to be “reachable” in S if it can be constructed by playing the game from an empty board with some sequence of pieces. More precisely, a state x is “reachable” if there exists a word

$$\sigma_1\sigma_2 \dots \sigma_i \in S$$

such that

$$x = e \cdot \sigma_1\sigma_2 \dots \sigma_i.$$

The semigroup is precisely the set of transformations S given by concatenating pieces and possible positions for those pieces on the board (p, ξ) , along with the set of game states reachable from the empty board using those $(p, \xi) \in S$.

Definition 1 (X, S) is a finite transformation semigroup, which we will call the **Tetris semigroup** of P on the board with dimensions $n \times k$.

3 Analysis

In order to obtain a full description of (X, S) in terms of transformations, we implemented the rules of Tetris in Python. Here, we will consider the complexity of a few variants of the game. Computation can be done on the 3×3 and 3×4 size game-board with tri-ominoes as described below, but any larger is currently out of reach of the computational capabilities of the GAP algebra system and SgpDec.

3.1 Tri-Tris

Standard Tetris has an extremely large state-space. Germundsson [8] estimates that it is on the order of 2^{200} , this estimate is corroborated by the later constructibility result of [10]. Therefore we will consider a variant of Tetris on an $n \times k$ board, using *triominoes* (Fig. 2) rather than tetriminoes.

Let $P = \{\text{LS, RS, LUS, RUS, V, H}\}$, then the corresponding game (and semigroup) we will call *Tri-tris* accordingly.

Fig. 2 Triominoes and corresponding labels used in the implementation

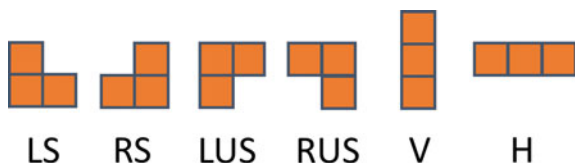


Table 1 Bounding the aperiodic complexity of Tri-tris on boards of different size by the length of the longest proper subduction chain, $h_s(X)$. All boards with width $n = 3$ have 11 generators. When n is increased to 4 the semigroup is too large for GAP to handle

Board dimensions	$ X $	$ S $	$h_s(X)$
3×3	35	2,056	13
3×4	135	259,726	32
3×5	709	–	–

We will mostly consider the setting where $n = 3$. In this case, the horizontal line triomino H is the identity, except when there are cells in the top line, in which case it maps to E . We will exclude H as it doesn't add significantly to the system.

3.2 Aperiodicity

A transformation semigroup is called *aperiodic* if all of its subgroups are trivial. We have found Tetris to be aperiodic for $n \leq 3$ and $k \leq 5$. We pose this as an open problem for larger board sizes.

Open Problem 1 (Is Tetris aperiodic?) *For any $\sigma \in S$ does there exist a $k > 0$ such that, for all $x \in X$, $x \cdot \sigma^{k+1} = x \cdot \sigma^k$. If so, the Tetris semigroup (X, S) is aperiodic, and the corresponding KR decomposition contains no nontrivial permutation groups.*

The aperiodic complexity of a transformation semigroup is the least number of identity-reset components (i.e., direct products of flip-flops) that must be wreathed together to emulate it (Table 1).

If Tetris is always aperiodic, this means that there are no internal symmetries for the holonomy decomposition to expose. The complexity increases extremely quickly as the board size increases. In the next section, we will consider a rule modification that introduces these symmetries.

4 Periodic Tri-Tris

The most straightforward rule modification to Tetris that gives the system reversibility (groups in the holonomy decomposition) is replacing the end state E with the empty board e . In this version of the game, any move that would previously have caused a loss, now returns the game to the empty board (Table 2).

Table 2 The complexity of periodic Tri-tris on boards of different sizes, showing the groups present in the holonomy decomposition given by SgpDec. Replacing the losing state E with the empty board state e , the semigroups become much larger. A computation for a 3×4 board with a reduced generator set is included as well

Board dimensions	$ X $	$ S $	Holonomy groups present
3×3	34	118,637	$(4, C_2 \times C_2), (3, S_3), (2, C_2)$
3×4	135	–	–
$3 \times 4,$ $P =$ $\{RS, LUS, RUS, V\}$	116	–	$(4, C_2), (5, S_5), (4, S_4), (3, S_3), (2, C_2)$

4.1 Periodic Tri-Tris: 3×3 Case

Non-trivial holonomy permutation groups appear in this case. Consider the empty state together with the three other states: {empty, 6, 12, 26}, whose non-empty states are visualized pictorially in Fig. 3. Figure 4 shows how the members of the holonomy group $C_2 \times C_2$ act on tiles of this set. We can find pictorial representations of some of the states and the transformations to more easily visualize what this figure is describing within the games (Fig. 3). Generally, the way that these groups permute the tile is that the words “reset” the state by exceeding the length of the board, and then construct the new state. The non-abelian group S_3 also appears in the holonomy decomposition.

4.2 Periodic Tri-Tris: 3×4 Case with Reduced Generating Set

If we increase the board size to 3×4 , and let $P = \{RS, LUS, RUS, V\}$, we see that the holonomy decomposition contains the full symmetric group S_5 , acting on the set

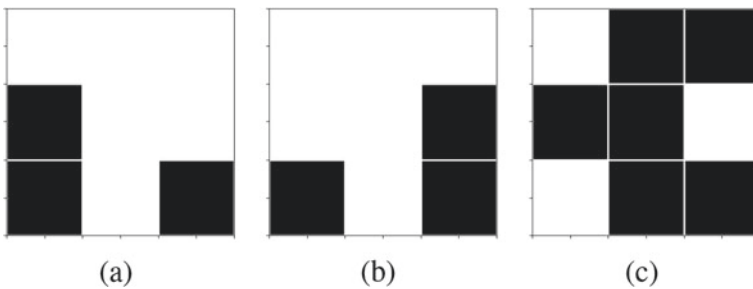


Fig. 3 Visualizations of the states in Fig. 4. **a** State 6, **b** State 12, **c** State 26

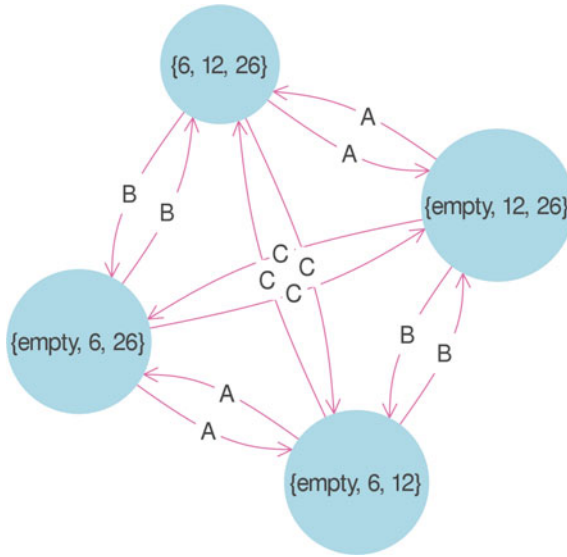


Fig. 4 Holonomy transformation group $(4, C_2 \times C_2)$ where $A = V_0LS_1V_2V_0RS_1V_1V_2V_0RS_1V_1$, $B = V_0LS_1V_2V_1V_0V_2LS_0V_1V_0V_2LS_0V_1$, $C = V_0V_2LS_0V_1$

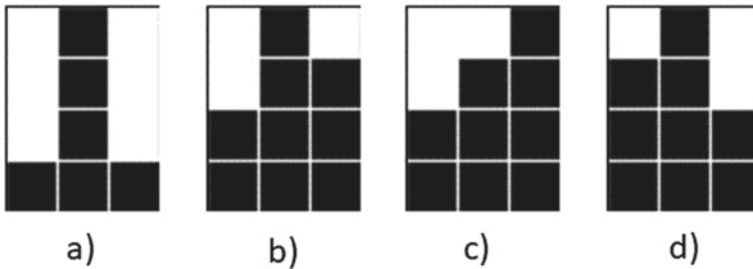


Fig. 5 Visualizations of the states permuted by S_5 , in addition to the empty state, in the Tetris semi-group with $n \times k = 3 \times 4$, $P = \{RS, LUS, RUS, V\}$. **a** State 4, **b** State 11, **c** State 13, **d** State 16

$Z = \{\text{empty}, 4, 11, 13, 16\}$, illustrated in Fig. 5. Although this group is too large to visualize in the same fashion as $C_2 \times C_2$, we can describe some generators of the permutator group. For instance, the word in Eq. 1 is a 5-cycle on Z , and the word in Eq. 2 is a 2-cycle on $(11, 16)$ and a 3-cycle on $(\text{empty}, 4, 13)$.

$$\begin{aligned}
& V_1 L U S_0 V_2 V_0 R S_1 R S_1 L U S_0 R S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 V_0 V_2 V_1 V_2 L U S_0 R S_1 \\
& V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 V_0 V_2 V_1 V_2 L U S_0 R S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 V_0 V_2 V_1 V_2 \\
& L U S_0 R S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 V_0 V_2 V_1 V_2 \\
& L U S_0 R S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 V_0 V_2 V_1 V_2 L U S_0 R S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 \\
& = (\text{empty}, 4, 11, 13, 16) \tag{1}
\end{aligned}$$

$$\begin{aligned}
& V_1 L U S_0 V_2 V_0 R S_1 R S_1 V_0 V_1 R S_0 V_2 V_0 V_2 L U S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 R U S_1 \\
& V_0 V_2 L U S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 R U S_1 V_0 V_2 L U S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 V_1 \\
& R U S_1 V_0 V_2 L U S_1 V_0 V_1 V_2 R S_0 R S_0 V_2 V_0 = (\text{empty}, 4, 13)(11, 16) \tag{2}
\end{aligned}$$

These two permutations generate the group S_5 .

5 SNAGs and Computation in Tetris

The appearance of the symmetric group S_5 in the transformation semigroup of 3×4 periodic Tri-tris shows that the smallest simple nonabelian group A_5 (the alternating group of 60 even permutations on 5 elements) can be emulated by this Tetris semigroup. Ensemble techniques of Nehaniv et al. [15] for computing with finite simple nonabelian groups (SNAGs) now entail that periodic Tri-tris is capable of *finitary universal computation*. This means every function $f : X \rightarrow Y$ for any finite sets X and Y can be realized via an implementation using an encoding into parallel running copies of this Tetris game.

Theorem 1 *The periodic 3×4 Tri-tris game is finitarily computationally universal.*

Sketch of Proof (construction of [15]). Let $n = \lceil \log_{60} |X| \rceil$ and $m = \lceil \log_{60} |Y| \rceil$. For each permutation π of Z in A_5 , fix a particular sequence w_π of Tetris events yielding π . One encodes distinct members of X each uniquely into n such sequences $(w_{\pi_1}, \dots, w_{\pi_n})$. Similarly, encode members of Y uniquely in m -tuples of permutations (π'_1, \dots, π'_m) , $\pi'_i \in A_5$. This yields an encoding of f as a mapping from n -tuples of the 60 different w_π sequences to m -tuples of permutations in A_5 . Now by a theorem of Maurer and Rhodes [14], each of the m components of the encoded f can be computed by some fixed *polynomial expression* over this SNAG. That is, each is some finite concatenation of the fixed sequences w_π giving permutations in A_5 and n free variables which take values in event sequences according to the encoding of X (with repetitions possible). The evaluation of these polynomial expressions with sequences encoding a member of X substituted in for the n variables consists of running Tri-tris and permutes states in parallel copies of the game (each in a configuration from Z). The result in Tri-tris comprises m permutations of Z lying in A_5 uniquely encoding the value of f in Y . It suffices to use $5m$ copies of Tri-tris since $|Z| = 5$ to determine the m permutations encoding $y = f(x)$ with $x \in X$, $y \in Y$; actually since we are dealing with permutations $4m$ copies of the game suffice.

6 Conclusion

We cast Tetris as a finite transformation semigroup, and show that the complexity of the game grows very quickly with the size of the game board. Modifying the rules of Tetris to restart on completion yields finite simple nonabelian groups (SNAGs) in the holonomy decomposition. This entails finite universal computational capacity of periodic variants of Tetris. While we found computationally that non-periodic Tetris examples had only trivial subgroups in their decompositions, it remains an open problem whether this is the case in all non-periodic variants. It also remains to determine the Krohn-Rhodes complexity and which SNAGs occur in other periodic versions of Tetris.

Acknowledgements We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2019-04669. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), numéro de référence RGPIN-2019-04669.

Appendix: Krohn-Rhodes Theory and the Holonomy Decomposition

The Krohn-Rhodes (KR) theorem describes a general decomposition of transformation semigroups in terms of wreath products of the finite simple groups and the flip-flop monoid. A visualization of the flip-flop monoid is shown in Fig. 6.

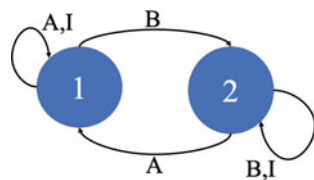
Theorem 2 (Krohn-Rhodes decomposition [12]) *A finite transformation semigroup (X, S) , with states X and semigroup S acting on states by transformations, has a decomposition*

$$(X, S) \text{ divides } H_1 \wr H_2 \wr H_3 \dots \wr H_n$$

with components H_1, H_2, \dots, H_n , such that each H_i is a finite simple group dividing S or the flip-flop monoid.

The decomposition given by this theorem tends to be far from optimal in practice. Therefore, most practical implementations of semigroup decomposition use the holonomy method described in [7].

Fig. 6 The flip-flop monoid on $X = \{1, 2\}$ is given by the set of transformations $S = \{A, B, I\}$



We will reproduce the relevant definitions and theorems here. Define the set Q as

$$Q = \{X \cdot s \mid s \in S\} \cup \{X\} \cup \{\{a\} \mid a \in X\}$$

then we can define a relation on Q called subduction.

Definition 2 (*Subduction*) Let S^I denote S with a new identity element appended. Given, $A, B \in Q$, we define an reflexive, transitive relation on Q ,

$$A \leq B \iff \exists s \in S^I, A \subseteq B \cdot s$$

Furthermore, let $A < B$ if $A \leq B$ but not $B \leq A$. This relation, which we will call *subduction*, induces an equivalence relation on Q : $A \equiv B \iff A \leq B$ and $B \leq A$. For each equivalence class A/\equiv in Q/\equiv , let \bar{A} be a unique representative.

Definition 3 (*Tiles*) Define A to be a *tile* of B if $A \subsetneq B$ and

$$\forall Z \in Q, (A \leq Z \leq B \implies Z = A \text{ or } Z = B)$$

If $A \in Q$ with $|A| > 1$, the set of tiles of A is $\Theta_A \subset Q$

Definition 4 (*Holonomy group*) The *holonomy group*, written H_A , of A is the set of permutations of Θ_A induced by the elements of S^I . If we let H_A act on Θ_A , then (Θ_A, H_A) is the holonomy permutation group of A .

Definition 5 (*Height of an Image Set*) The *height* of $A \in Q$ is $h(A)$, where $h(A)$ is the length of the longest strict subduction chain up to A .

We are now able to state the holonomy decomposition theorem, which asserts that the semigroup (X, S) divides a cascade product, from which the Krohn-Rhodes (KR) decomposition (Theorem 2) can be derived. The holonomy theorem describes the transformation semigroup (X, S) in terms of symmetries in the way transformations in S act on the set of tiles of the $\bar{A} \in Q$.

Theorem 3 (Holonomy decomposition [7]) *Let (X, S) be a finite transformation semigroup, with $h = h(X)$ the height of X . For each i with $1 \leq i \leq h$, let*

$$(\Phi_i, \mathfrak{H}_i) = \prod_{\{A \in Q : h(A)=i, \bar{A}=A\}} (\Theta_{\bar{A}}, H_{\bar{A}})$$

(Φ_i, \mathfrak{H}_i) is a permutation group and $(\Phi_i, \overline{\mathfrak{H}}_i)$ is the permutation-reset transformation semigroup obtained by appending all constant maps to \mathfrak{H}_i . Then

$$(X, S) \text{ divides } (\Phi_1, \overline{\mathfrak{H}}_1) \wr (\Phi_2, \overline{\mathfrak{H}}_2) \wr \cdots \wr (\Phi_h, \overline{\mathfrak{H}}_h).$$

References

1. Baccherini, D., Merlini, D.: Combinatorial analysis of Tetris-like games. *Discrete Math.* **308**(18), 4165–4176 (2008)
2. Brzustowski, J.: Can you win at Tetris? Master’s thesis, University of British Columbia (1992)
3. Burgiel, H.: How to lose at Tetris. *Math. Gaz.* **81**(491), 194–200 (1997)
4. Demaine, E., Hohenberger, S., Liben-Nowell, D.: Tetris is hard, even to approximate. In: *Computing and Combinatorics*, pp. 351–363. LNCS, vol. 2697. Springer (2003)
5. Egri-Nagy, A., Mitchell, J.D., Nehaniv, C.L.: SgpDec: cascade (de)compositions of finite transformation semigroups and permutation groups. In: *International Congress on Mathematical Software*, pp. 75–82. LNCS, vol. 8592. Springer (2014)
6. Egri-Nagy, A., Nehaniv, C.L.: Ideas of the holonomy decomposition of finite transformation semigroups. *RIMS Kôkyûroku* **2051**, 43–45 (2017)
7. Eilenberg, S.: *Automata, Languages, and Machines*, Vol. B. Academic Press (1976)
8. Germundsson, R.: *A Tetris Controller: An Example of a Discrete Event Dynamic System*. Linköping University, Sweden (1991)
9. Hoad, P.: Tetris: how we made the addictive computer game (2014). <https://www.theguardian.com/culture/2014/jun/02/how-we-made-tetris>
10. Hoogeboom, H.J., Kosters, W.A.: The Theory of Tetris. *Nieuwsbr. Ned. Ver. voor Theoret. Inform.* **9**, 14–21 (2005)
11. Jim Pattison Group: *Guinness World Records: Gamer’s edition* (2011)
12. Krohn, K., Rhodes, J.: Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Trans. Am. Math. Soc.* **116**, 450–464 (1965)
13. Maler, O.: On the Krohn-Rhodes cascaded decomposition theorem. In: *Time for Verification: Essays in Memory of Amir Pnueli*, pp. 260–278, LNCS, vol. 6200. Springer (2010)
14. Maurer, W.D., Rhodes, J.L.: A property of finite simple non-abelian groups. *Proc. Am. Math. Soc.* **16**(3), 552–554 (1965)
15. Nehaniv, C.L., Rhodes, J., Egri-Nagy, A., Dini, P., Rothstein Morris, E., Horváth, G., Karimi, F., Schreckling, D., Schilstra, M.J.: Symmetry structure in discrete models of biochemical systems: natural subsystems and the weak control hierarchy in a new model of computation driven by interactions. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **373**(2046), 20140,223 (2015)
16. Weisstein, E.W.: Tetromino – from Wolfram MathWorld (2003). <https://mathworld.wolfram.com/Tetromino.html>

Differential Equations Using Generalized Derivatives on Fractals



Herb Kunze, Davide La Torre, Franklin Mendivil, and Edward R. Vrscay

Abstract In a previous paper [11] we introduced the notion of a μ -derivative and showed how to formulate differential equations in terms of this derivative. In this paper, we extend this approach to the definition of a weak derivative which provides a framework for solving variational problems with respect to fractal measures. We apply our method to a specific boundary value problem, namely a 1D eigenvalue problem over a fractal measure.

Keywords Fractals · Fractal measure · Derivatives · Weak derivatives

1 Introduction: Derivatives with Respect to a Fractal Measure

In this paper we present a framework for solving variational problems with respect to a fractal measure by extending the ideas from [11]. Our theory uses the weak formulation and thus we define the weak derivative and the resulting Sobolev spaces in the natural way. For the one-dimensional problems we discuss in this paper, the variational problems can be transformed by an appropriate change-of-variable into

H. Kunze (✉)
University of Guelph, Guelph, Canada
e-mail: hkunze@uoguelph.ca

D. La Torre
SKEMA Business School, Université de la Côte-d'Azur, Sophia-Antipolis Campus,
Valbonne, France
e-mail: davide.latorre@skema.edu

F. Mendivil
Acadia University, Wolfville, Canada
e-mail: franklin.mendivil@acadiau.ca

E. R. Vrscay
University of Waterloo, Waterloo, Canada
e-mail: ervrscay@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_8

a problem involving Lebesgue measure and thus many of the classical results can be used directly. Problems in higher dimensions require a substantial reworking of the classical theory and are the subject of a future paper in preparation.

In a previous paper [11] we introduced the notion of a μ -derivative and we discussed how to formulate differential equations in which the derivative is replaced by a μ -derivative. We considered the equivalent integral equation,

$$u(x) = u_0 + \int_0^x f(t, u(t)) d\mu(t), \quad (1)$$

where μ is a fractal (Borel probability) measure, assumed to be nonatomic, on $[0, 1]$. We studied the existence and uniqueness of solutions to these fractal integral equations based on the Picard operator. Our main interest was in the fractal nature of the solutions and we used Iterated Function Systems (IFS) to investigate the behaviour and self-similarity of these solutions. As usual we can try to formulate an integral equation into an equivalent differential form. Motivated by this we defined the μ -derivatives of a function G to be

$$D_{\mu}^{+}(G)(x) := \lim_{h \rightarrow 0^{+}} \frac{G(x+h) - G(x)}{\mu([x, x+h])}.$$

In a similar way, we can define

$$D_{\mu}^{-}(G)(x) := \lim_{h \rightarrow 0^{+}} \frac{G(x) - G(x-h)}{\mu([x-h, x])}.$$

Whenever the two limits are equal we label their common value $D_{\mu}(G)(x)$ and say that G is μ -differentiable at x [11].

A version of the Fundamental Theorem of Calculus holds [11] so that the integral equation (1) becomes the μ -differential initial value problem,

$$D_{\mu}(u)(x) = f(x, u(x)), \quad u(0) = u_0. \quad (2)$$

The following results are very useful when dealing with calculations involving the notion of μ -derivative.

Proposition 1 ([11]) *Let us suppose that μ is non-atomic and let $F : K = \text{supp}(\mu) \rightarrow [0, 1]$ be the cumulative of μ and $F^{-1} : [0, 1] \rightarrow K$ be its inverse. Then, given a function $f : K \rightarrow \mathbb{R}$, the following change of variable rule holds:*

$$\int_K f(x) d\mu(x) = \int_0^1 f(F^{-1}(x)) dx, \quad (3)$$

where dx indicates integration over Lebesgue measure on $[0, 1]$.

Proposition 2 ([11]) *Let us suppose that μ is non-atomic and let $F : K = \text{supp}(\mu) \rightarrow [0, 1]$ be the cumulative of μ and $F^{-1} : [0, 1] \rightarrow K$ be its inverse. Then,*

given a function $f : K \rightarrow \mathbb{R}$, the following chain rule holds

$$D_\mu f(y) = \frac{d}{dx} f(F^{-1}(x))|_{x=F(y)}, \text{ for } \mu\text{-a.e. } y, \tag{4}$$

where dx denotes Lebesgue measure and $y = F^{-1}(x)$. Moreover, the following formula for higher-order derivatives holds:

$$D_\mu^n f(y) = \frac{d^n}{dx^n} f(F^{-1}(x))|_{x=F(y)}. \tag{5}$$

Using these properties it is not hard to show that the following version of integration by parts holds (where $\text{cov}(A)$ is the convex hull of A).

Proposition 3 *Let us suppose $[a, b] = \text{cov}(\text{supp}(\mu))$. Then the following formula holds:*

$$\int_a^b D_\mu f(t) g(t) d\mu(t) = f(b)g(b) - f(a)g(a) - \int_a^b f(t)D_\mu g(t) d\mu(t). \tag{6}$$

In the next sections we extend this approach to deal with boundary value problems (BVP) and with particular application to a simple example. We then introduce the notion of a weak μ -derivative and present a variational formulation of the BVP.

The paper is organized as follows. Section 2 presents the notion of μ -weak derivative and the definition of the Hilbert space $H_\mu^1(K)$ along with an application to a one-dimensional eigenvalue BVP. Section 3 recalls the basic definitions of Iterated Function Systems and the notion of attractor. Section 4 presents some convergence results and Sect. 5 contains some concluding remarks.

We provide a brief excursion into this topic with the intention to interest the reader in the possibilities. Because of space limitations we do not provide proofs. For a much more in-depth discussion, including proofs and extensions we invite the reader to read our forthcoming paper (in preparation).

It is important to mention that our work here (as in [11]) is strongly related to other previous work in analysis on *time-scales* (see [2, 8] and the references therein), in *measure differential equations* (see [3, 17] and the references therein) and also in Stieltjes derivatives (as nicely explained in [16]). More recent work in time-scale analysis which is strongly related to the current paper can be found in [4] (and its references). The papers [13, 14] present another method for defining calculus on subsets of $E \subset \mathbb{R}$ which is geometrically defined and intrinsic to E (and so do not depend on the existence of a measure on E). Their results imply the results in [11] in the case of a “uniform” measure on E and thus could be used as an alternative approach to ours.

From the perspective of applications, the use of fractal derivatives in physics has been recognized, for example, in [6] as have been variational methods [7]. There is an enormous literature on the subject which this paper cannot hope to address even in part. Here we simply mention [5, 10, 18] as noteworthy contributions to the field.

2 The Weak Formulation and $H_\mu^1(K)$

Let $K \subset \mathbb{R}$ be a given compact “fractal” set with convex hull $\text{cov}(K) = [a, b]$ and μ be a Borel probability measure supported on it. For a given function $\phi : K \rightarrow \mathbb{R}$, we denote by $D_\mu \phi$ the μ -derivative of ϕ with respect to μ at x (which is well-defined at μ -almost all x). In the sequel we denote by $C_c^1(K)$ the set of all functions for which the μ -derivative exists, it is continuous and $\phi(a) = \phi(b) = 0$. Given a function $u : K \rightarrow \mathbb{R}$, the *weak μ -derivative* of u is a function $g : K \rightarrow \mathbb{R}$ which satisfies

$$\int_K u D_\mu \phi d\mu = - \int_K g \phi d\mu \text{ for all } \phi \in C_c^1(K). \quad (7)$$

We also denote the weak μ -derivative of u by $D_\mu u$. Using the fact the μ is supported on K , the previous integral can be rewritten as

$$\int_{[a,b]} u D_\mu \phi d\mu = - \int_{[a,b]} g \phi d\mu \quad (8)$$

where μ also denotes its extension to $[a, b]$ (i.e., the measure $\mu(A) = \mu(K \cap A)$).

As usual, we define the space $L_\mu^p(K)$ as the set of all functions $u : K \rightarrow \mathbb{R}$ that satisfy the condition,

$$\int_K |u|^p d\mu < +\infty. \quad (9)$$

In a similar way, we denote by $W_\mu^{1,p}(K)$ the following set,

$$W_\mu^{1,p}(K) = \{u : K \rightarrow \mathbb{R}, u \in L_\mu^p(K) : D_\mu u \text{ exists and } D_\mu u \in L_\mu^p(K)\}, \quad (10)$$

with $H_\mu^1(K) = W_\mu^{1,2}(K)$. It is not complicated to prove that the space $H_\mu^1(K)$ is Hilbert with respect to the inner product,

$$\langle u, v \rangle = \int_K D_\mu u(x) D_\mu v(x) d\mu + \int_K u(x) v(x) d\mu,$$

and induced norm

$$\|u - v\|_{H_\mu^1} = \|D_\mu u - D_\mu v\|_{L_\mu^2(K)} + \|u - v\|_{L_\mu^2(K)}.$$

Example: We now consider the Dirichlet problem taking the form,

$$D_\mu^2 u(x) + \lambda u(x) = f(x), \quad u(0) = 0, \quad u(1) = 0. \quad (11)$$

Following the standard procedure, we obtain an equivalent formulation by first multiplying both sides by a test function $\xi \in C_c^1(K)$. Integration by parts leads to

$$\begin{aligned}
 \int_K f(x)\xi(x)d\mu(x) &= \int_K D_\mu^2 u(x)\xi(x)d\mu(x) + \lambda \int_K u(x)\xi(x)d\mu(x) \\
 &= D_\mu u(b)\xi(b) - D_\mu u(a)\xi(a) + \int_K D_\mu u(x)D_\mu \xi(x)d\mu(x) \\
 &\quad + \lambda \int_K u(x)\xi(x)d\mu(x) \\
 &= \int_K D_\mu u(x)D_\mu \xi(x)d\mu(x) + \lambda \int_K u(x)\xi(x)d\mu(x).
 \end{aligned}$$

We arrive at the variational form,

$$\int_K D_\mu u D_\mu \xi \, d\mu + \lambda \int_K u \xi \, d\mu = \int_K f \xi \, d\mu. \tag{12}$$

If we define the bilinear form,

$$b(u, v) := \int_K D_\mu u(x)D_\mu v(x)d\mu(x) + \lambda \int_K u(x)v(x)d\mu(x), \tag{13}$$

and the linear form,

$$\theta(v) = \int_K f(x)v(x)d\mu(x), \tag{14}$$

then (11) can be written as follows: Find $u \in H_\mu^1(K)$ such that

$$b(u, v) = \theta(v) \tag{15}$$

for any $v \in H_\mu^1(K)$. The existence and uniqueness of solutions to (15) can be proved using the classical Lax-Milgram Theorem.

We conclude this section by showing how our method in [11] may be used to the 1D eigenvalue-BVP in Eq. (11). Once again assuming that μ is non-atomic, we define the variable $t = F(x) = \mu((-\infty, x])$, where $F(x)$ denotes the cumulative distribution function associated with μ . Also let $x = F^{-1}(t)$. Using the change of variable presented in Proposition 2, we obtain

$$D_\mu^2 u(x) + \lambda u(x) = \frac{d^2}{dt^2} u(F^{-1}(t))|_{t=F(x)} + \lambda u(x) = 0, \quad u(a) = 0, \quad u(b) = 0,$$

which is equivalent to

$$\frac{d^2}{dt^2} u(F^{-1}(t))(t) + \lambda u(F^{-1}(t)) = 0.$$

By defining $g(t) = u(F^{-1}(t))$, this can be written as

$$\frac{d^2}{dt^2}g(t) + \lambda g(t) = 0, \quad g(0) = 0, \quad g(1) = 0.$$

This, of course, is the classical “vibrating string” eigenvalue problem on $[0, 1]$ with solutions $\lambda_n = (n\pi)^2$ and $g_n(t) = \sin(n\pi t), n \geq 1$. From these, the solutions to (11) may be expressed in terms of the variable x as simply $u_n(x) = \sin(n\pi F(x))$.

In each of Figs. 1, 2 and 3 are shown histogram approximations to the invariant measure μ and its CDF function F_μ along with the first three eigenfunctions $u_n(x)$. In Fig. 1, the IFS is $w_1(x) = x/3$ and $w_2(x) = x/3 + 2/3$ with probabilities $p_1 = p_2 = 1/2$. This IFSP generates a “uniform” distribution on the classical middle-1/3 Cantor set. The same two IFS maps are employed in Fig. 2, but with probabilities $p_1 = 2/5$ and $p_2 = 3/5$. The larger weight “towards the right” is evident in all portions of μ, F_μ (its CDF) and the eigenfunctions. In Fig. 3, the two IFS maps are $w_1(x) = x/2$ and $w_2(x) = x/2 + 1/2$ with probabilities $p_1 = 2/5$ and $p_2 = 3/5$. Here, the attractor is $[0, 1]$. Once again, the unequal weighting produces a (self-similar) “shift” of the measure to the right.

Note that in both Figs. 1 and 2 the eigenfunctions are illustrated by extending them to be constant over the “gaps” in the complement of the Cantor set. (These functions are supported only on the Cantor set itself.) This is done in order to make their graphs visible.

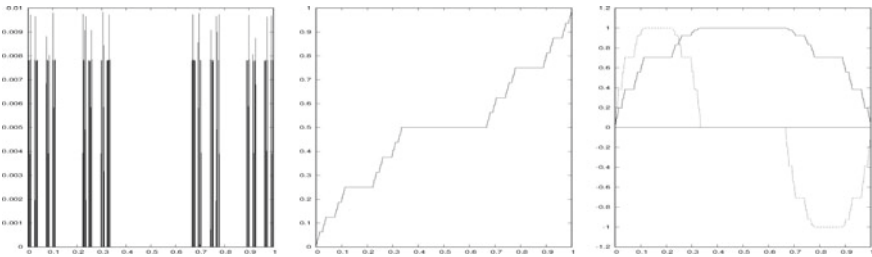


Fig. 1 “Uniform Cantor measure” μ , CDF F_μ , and first three eigenfunctions $u_n(x) = \sin(n\pi F_\mu(x))$

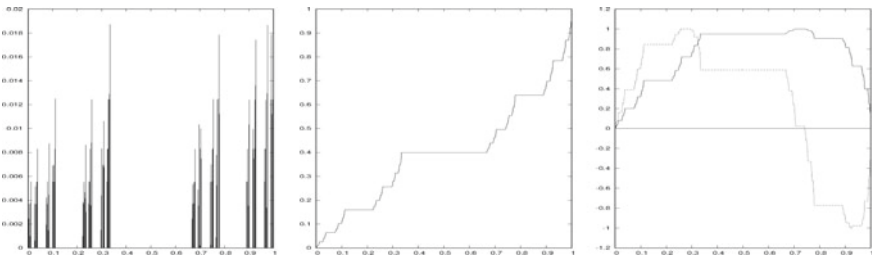


Fig. 2 “Non-uniform Cantor measure” μ , CDF F_μ , and first three eigenfunctions $\sin(n\pi F_\mu(x))$

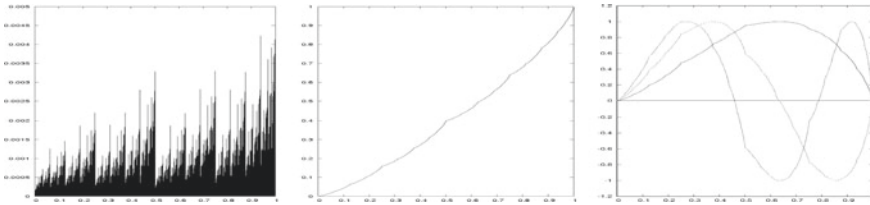


Fig. 3 Non-uniform fully supported measure μ , CDF F_μ , and first three eigenfunctions $\sin(n\pi F_\mu(x))$

3 Some Basics of Iterated Function Systems

In what follows, we let (X, d) denote a compact metric space. An N -map *Iterated Function System* (IFS) on X , $\mathbf{w} = \{w_1, \dots, w_N\}$, is a set of N contraction mappings on X , i.e., $w_i : X \rightarrow X, i = 1, \dots, N$, with contraction factors $c_i \in [0, 1)$. (See [1, 9, 12].) Associated with an N -map IFS is the following set-valued mapping $\hat{\mathbf{w}}$ on the space $\mathcal{H}(X)$ of nonempty compact subsets of X ,

$$\hat{\mathbf{w}}(S) := \bigcup_{i=1}^N w_i(S), \quad S \in \mathcal{H}(X). \tag{16}$$

Theorem 1 ([9]) For $A, B \in \mathcal{H}(X)$,

$$h(\hat{\mathbf{w}}(A), \hat{\mathbf{w}}(B)) \leq ch(A, B) \quad \text{where } c = \max_{1 \leq i \leq N} c_i < 1 \tag{17}$$

and h denotes the Hausdorff metric on $\mathcal{H}(X)$.

Corollary 1 ([9]) There exists a unique set $A \in \mathcal{H}(X)$, the attractor of the IFS \mathbf{w} , such that

$$A = \hat{\mathbf{w}}(A) = \bigcup_{i=1}^N w_i(A). \tag{18}$$

Moreover, for any $B \in \mathcal{H}(X)$, $h(A, \hat{\mathbf{w}}^n B) \rightarrow 0$ as $n \rightarrow \infty$.

An N -map *Iterated Function System with Probabilities (IFSP)* (\mathbf{w}, \mathbf{p}) is an N -map IFS \mathbf{w} with associated probabilities $\mathbf{p} = \{p_1, \dots, p_N\}$, $\sum_{i=1}^N p_i = 1$. Let $\mathcal{M}(X)$ denote the set of probability measures on (Borel subsets of) X with Monge-Kantorovich distance d_{MK} : For $\mu, \nu \in \mathcal{M}(X)$,

$$d_{MK}(\mu, \nu) = \sup_{f \in Lip_1(X)} \left[\int f d\mu - \int f d\nu \right], \tag{19}$$

where $Lip_1(X) = \{f : X \rightarrow \mathbb{R} \mid |f(x) - f(y)| \leq d(x, y)\}$. The metric space $(\mathcal{M}(X), d_{MK})$ is complete [1, 9].

Associated with an N -map IFSP is a mapping $M : \mathcal{M} \rightarrow \mathcal{M}$, often referred to as the *Markov operator*, defined as follows. Let $\nu = M\mu$ for any $\mu \in \mathcal{M}(X)$. Then for any measurable set $S \subset X$,

$$\nu(S) = (M\mu)(S) = \sum_{i=1}^N p_i \mu(w_i^{-1}(S)). \quad (20)$$

Theorem 2 ([9]) For $\mu, \nu \in \mathcal{M}(X)$,

$$d_{MK}(M\mu, M\nu) \leq c d_{MK}(\mu, \nu). \quad (21)$$

Corollary 2 ([9]) There exists a unique measure $\bar{\nu} \in \mathcal{M}(X)$, the invariant measure of the IFSP (\mathbf{w}, \mathbf{p}) , such that

$$\bar{\mu}(S) = (M\bar{\mu})(S) = \sum_{i=1}^N p_i \bar{\mu}(w_i^{-1}(S)). \quad (22)$$

Moreover, for any $\nu \in \mathcal{M}(X)$, $d_{MK}(\bar{\mu}, M^n \nu) \rightarrow 0$ as $n \rightarrow \infty$.

Theorem 3 ([9]) The support of the invariant measure $\bar{\mu}$ of an N -map IFSP (\mathbf{w}, \mathbf{p}) is the attractor A of the IFS \mathbf{w} , i.e., $\text{supp } \bar{\mu} = A$.

The next result is rather technical but is used in our convergence results in Sect. 4. The proof uses the fact that an IFSP on \mathbb{R} induces a natural IFS-type operator on cumulative distribution functions which is contractive in the uniform norm.

Theorem 4 Let (\mathbf{w}, \mathbf{p}) be an N -map IFSP with non-atomic invariant measure μ . Let $[a, b] = \text{cov}(\text{supp}(\mu))$ and suppose that $w_i([a, b]) \cap w_j([a, b])$ for $i \neq j$ either empty or consisting of one point.

Let μ_0 be any initial Borel probability measure supported on $[a, b]$, $\mu_{n+1} = M\mu_n$, $F : [a, b] \rightarrow [0, 1]$ be defined as $F(x) = \mu([a, x])$ and $F_n : [a, b] \rightarrow [0, 1]$ be defined as $F_n(x) = \mu_n([a, x])$. Then $F_n \rightarrow F$ uniformly on $[a, b]$.

4 Convergence of Solutions

We now discuss a simple convergence result for the above eigenvalue problem. We restrict our presentation to the simplest case for clarity and brevity; more general results are certainly possible (including results on the variational problem (12)) but we leave them to our future paper.

Take an IFSP (\mathbf{w}, \mathbf{p}) and initial measure μ_0 so that they both satisfy the conditions of Theorem 4 and consider the sequence of eigenvalue problems: Find $u \in H^1_{\mu_n}([a, b])$ so that

$$\int_{[a,b]} D_{\mu_n} u D_{\mu_n} v d\mu_n + \lambda \int_{[a,b]} uv d\mu_n = 0, \quad \text{for all } v \in H^1_{\mu_n}([a, b]). \quad (23)$$

Proposition 4 *Given μ_n and u_n as above we have that $u_n \rightarrow u$ uniformly and u is solution to the problem:*

$$\int_{[a,b]} D_{\mu} u D_{\mu} v d\mu + \lambda \int_{[a,b]} uv d\mu = 0. \quad (24)$$

We end with a small taste of a more general problem. Start with μ_0 as the Lebesgue measure. Then the solutions u_n to the variational problems,

$$\int_{[a,b]} D_{\mu} u_n D_{\mu} v d\mu_n + \lambda \int_{[a,b]} u_n v d\mu_n = \int_{[a,b]} f v d\mu_n, \quad (25)$$

can be found by using more classical methods involving subproblems with weighted versions of the Lebesgue measure. Note that by using the definition of the Markov operator and a change of variable, the first term in (25) can be written as follows,

$$\begin{aligned} \int D_{\mu} u_n D_{\mu} v d\mu_n &= \int D_{\mu} u_n D_{\mu} v dM^n \mu_0 = \\ &= \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \int (D_{\mu} u_n D_{\mu} v) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0. \end{aligned}$$

Similarly,

$$\lambda \int uv d\mu_n = \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \lambda \int (uv) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0$$

and

$$\int f v d\mu_n = \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \int (f v) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0.$$

Thus the variational problem with respect to μ_n can be reformulated as follows: Find $u_n \in H^1_{\mu}([a, b])$ such that

$$\begin{aligned} & \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \int_{K_s} (D_\mu u_n D_\mu v) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0 + \\ & \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \lambda \int_{K_s} (uv) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0 = \\ & \sum_{\sigma_1, \dots, \sigma_n=1}^N p_{\sigma_1} p_{\sigma_2} \dots p_{\sigma_n} \int (fv) \circ w_{\sigma_1} \circ w_{\sigma_2} \circ \dots \circ w_{\sigma_n} d\mu_0. \end{aligned}$$

Notice that these integrals are all performed with respect to Lebesgue measure.

5 Conclusion

In [11] we introduced the notion of μ -derivative and we discussed how to formulate differential equations in which the derivative is replaced by a μ -derivative. In this paper, instead, we have extended this approach to the definition of weak derivative and to deal with boundary value problems. We have shown an application to a specific BVP, namely an eigenvalue problem, and presented a variational formulation of this problem in 1D. Future avenues include an extension of this approach to introduce weak partial derivatives to analyze variational problems on 2D fractals.

Acknowledgements This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) in the form of Discovery Grants (HK, FM and ERV).

References

1. Barnsley, M.F.: *Fractals Everywhere*. Academic Press, New York (1989)
2. Bohner, M., Peterson, A.: *Dynamic Equations on Time Scales*. Birkhäuser, Boston (2001)
3. Brogliato, B.: *Nonsmooth Mechanics. Models, Dynamics and Control*, 3rd ed. Springer-Verlag, Switzerland (2016)
4. Eckhardt, J., Teschl, G.: Sturm-Liouville operators on time scales. *J. Difference Equ. Appl.* **18**(11), 1875–1887 (2012)
5. Golmankhaneh, A.K., Tunc, C.: Stochastic differential equations on fractal sets. *Stochastics* (2019). <https://doi.org/10.1080/17442508.2019.1697268>
6. He, C.H., Shen, Y., Ji, F.Y., He, J.H.: Taylor series solution for fractal Bratu-type equation arising in electrospinning process. *Fractals* **28**(1), 205011 (2020)
7. He, J.H.: A fractal variational theory for one-dimensional compressible flow in a microgravity space. *Fractals*. <https://doi.org/10.1142/S0218348X20500243> (to appear)
8. Hilger, S.: Analysis on measure chains: a unified approach to continuous and discrete calculus. *Results Math.* **18**, 18–56 (1990)
9. Hutchinson, J.: Fractals and self-similarity. *Indiana Univ. J. Math.* **30**, 713–747 (1981)
10. Kesselböhmer, M., Samuel, T., Weyer, H.: A note on measure-geometric Laplacians. *Mon. Math.* **181**(3), 643–655

11. Kunze, H., La Torre, D., Mendivil, F., Vrscay, E.R.: Self-similarity of solutions to integral and differential equations with respect to a fractal measure. *Fractals* (2019). 1950014
12. Kunze, H., La Torre, D., Mendivil, F., Vrscay, E.R.: *Fractal-Based Methods in Analysis*. Springer, New York (2012)
13. Parvate, A., Gangal, A.D.: Calculus on fractal subsets of real line–I: Formulation. *Fractals* (1), 53–81 (2009)
14. Parvate, A., Gangal, A.D.: Calculus on fractal subsets of real line-II: Conjugacy with ordinary calculus. *Fractals* **19**(3), 271–290 (2011)
15. Petrovic, S.: Dynamic equations on the Cantor set. *PanAmer. Math. J.* **13**(4), 1–18 (2003)
16. Pouso, R.L., Rodríguez, A.: A new unification of continuous, discrete, and impulsive calculus through Stieltjes derivatives. *Real Anal. Exchange* **40**(2), 319–353 (2014/15)
17. Schmaedeke, W.W.: Optimal control theory for nonlinear vector differential equations containing measures. *SIAM J. Control* **3**(2), 231–280 (1965)
18. Wu, J., Wang, C.: Fractal Stokes’ theorem based on integration on fractal manifolds. *Fractals*. <https://doi.org/10.1142/S0218348X20500103>

Revisiting Path-Following to Solve the Generalized Nash Equilibrium Problem



Tangi Migot and Monica-G. Cojocaru

Abstract In this short paper, we present a generic path-following approach to tackle the generalized Nash equilibrium problem (GNEP) via its KKT conditions. This general formulation can be specialized to various smoothing techniques, including the popular interior-point method. We prove that under classical assumptions, there exists a path starting from an initial point and leading to an equilibrium of the GNEP. We also open the discussion on how one can derive numerical methods based on our approach.

Keywords Generalized Nash equilibrium problem · KKT conditions · GNEP · KKT · Interior-point method · Homotopy

Introduction

In the early 50s [30], Nash introduces a notion of equilibrium, the so-called Nash equilibrium, for non-cooperative N -player games where the payoff function of each player depends on the others' strategies. Later on, Arrow and Debreu [3] extended this notion to the generalized Nash equilibrium for games where both the payoff function and the set of feasible strategies depend on others' strategies. Initially motivated by economic applications, the notion of equilibrium in games has received a vivid interest thanks to its various applications in social science, biology, computer science or energy problems to cite few among others. These applications have motivated the evolution of the Nash equilibrium concept, and its use, to complex games that now require a deep understanding of theoretical and computational mathematics.

The study of numerical methods to compute one (or more) equilibrium of the generalized Nash equilibrium problem (GNEP) started two decades ago in the oper-

T. Migot (✉) · M.-G. Cojocaru
Department of Mathematics and Statistics, University of Guelph, Guelph N1G 2W1, Canada
e-mail: tmigot@uoguelph.ca

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

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_9

ational research community. Several approaches have been proposed in the literature: decomposition methods [14, 17, 18, 22, 27, 32], (quasi)-variational inequality type methods [26], penalty methods [15], Nikaido Isoda-function type methods [31, 33], ordinary differential equation based methods [6, 28], Newton type approaches [10, 25] and homotopy methods [11, 19]. We refer to [14, 22] for a review of the main numerical approaches and refer to [8, 29] for computing all the equilibria.

One approach to numerically tackle this problem is to replace each players' problem with its optimality conditions, the Karush-Kuhn-Tucker (KKT) conditions. Then, we obtain a non-linear equation formed by concatenating the N KKT systems, so-called GNEP KKT system. In [13], the authors introduced a potential reduction algorithm to solve this system. The method has later been extended to a hybrid potential reduction algorithm [10, 12] combining the original algorithm with the desired local convergence properties of the LP-Newton [21]. This approach has been shown to be successful from the practical point of view but also offers a global convergence under classical assumptions.

In this short paper, we present a generic homotopy method to tackle the GNEP KKT system. This general formulation allows us to then specialize this approach to various smoothing techniques, including the popular interior-point method. We prove that under classical assumptions, there exists a path starting from an initial point and leading to an equilibrium of the GNEP.

The paper is organized as follows. In Sect. 1, we introduce the necessary mathematical background and the GNEP. In Sect. 2, we present a homotopy method to solve the GNEP and prove that starting from an initial point, there exists a smooth path leading to a solution. In Sect. 3, we provide examples of homotopy, and, in particular, link the interior-point method with the GNEP. Finally, in Sect. 4, we discuss how this work can be extended to derive numerical methods.

1 The Generalized Nash Equilibrium Problem

The generalized Nash equilibrium problem (GNEP) is characterized by a set of N players, each of whom controls a finite set of variables $x^\nu \in \mathbb{R}^{n_\nu}$. We denote by $x := (x^1, \dots, x^N)^T$ the vector formed by all the decision variables, which has dimension $n := \sum_{\nu=1}^N n_\nu$. We denote by $x^{-\nu}$ the vector formed by all the players' decision variables except those of Player ν . We sometimes write $(x^\nu, x^{-\nu})$ instead of x , which does not mean that the block components of x are reordered. The goal of each player is to minimize their objective function $\theta_\nu(\cdot, x^{-\nu})$ subject to some constraint set $K_\nu(x^{-\nu})$. The key feature here is that each K_ν depends on variables beyond Player ν 's control. An equilibrium is any strategy x^* where no player can lower their objective function by unilaterally altering their strategy, i.e. for every $\nu \in \{1, \dots, N\}$ it holds

$$\theta_\nu(x^{*,\nu}, x^{*,-\nu}) \leq \theta_\nu(x^\nu, x^{*,-\nu}), \quad \forall x^\nu \in K_\nu(x^{*,-\nu}).$$

Stated otherwise, an N -player GNEP consists of N optimization problems with the player ν controlling n_ν variables and being subject to constraints:

$$\min_{x^\nu \in \mathbb{R}^{n_\nu}} \theta_\nu(x^\nu, x^{-\nu}) \text{ s.t. } x^\nu \in K_\nu(x^{-\nu}).$$

A typical choice of set K_ν is described by equality and inequality constraints, for instance

$$K_\nu(x^{-\nu}) := \{x^\nu \in \mathbb{R}^{n_\nu} : g^\nu(x^\nu, x^{-\nu}) \leq 0, h^\nu(x^\nu, x^{-\nu}) = 0\}, \quad (1)$$

where $g^\nu : \mathbb{R}^n \rightarrow \mathbb{R}^{m_\nu}$ and $h^\nu : \mathbb{R}^n \rightarrow \mathbb{R}^{p_\nu}$. We denote by $K(x) := \prod_{\nu=1}^N K_\nu(x^{-\nu})$.

Assume for all ν and all $x^{-\nu} \in \mathbb{R}^{-\nu}$ that $\theta_\nu(\cdot, x^{-\nu})$, $g^\nu(\cdot, x^{-\nu})$ are differentiable and convex, and, $h^\nu(\cdot, x^{-\nu})$ is affine. Moreover, assume that for all ν a constraint qualification holds for (1), then the GNEP is equivalent to the system formed by concatenating the KKT conditions of each optimization problem, [14]. In other words, x^* is a generalized Nash equilibrium if and only if for all ν there exists $(\lambda^\nu, \mu^\nu) \in \mathbb{R}^{m_\nu} \times \mathbb{R}^{p_\nu}$ such that the following system is satisfied:

$$\begin{aligned} \nabla_{x^\nu} \theta_\nu(x^*) + \nabla_{x^\nu} g^\nu(x^*)^T \lambda^\nu + \nabla_{x^\nu} h^\nu(x^*)^T \mu^\nu &= 0, \\ \lambda_i^\nu g_i^\nu(x^*) &= 0, \lambda_i^\nu \geq 0, g_i^\nu(x^*) \leq 0, \forall i = 1, \dots, m_\nu, \\ h_i^\nu(x^*) &= 0, \forall i = 1, \dots, p_\nu. \end{aligned} \quad (2)$$

Since the concatenation of the N system (2) is equivalent to the GNEP, we will use this system to solve the GNEP. This is a system of size $n + p + m$ where $n := \sum_{\nu=1}^N n_\nu$, $m := \sum_{\nu=1}^N m_\nu$, and $p := \sum_{\nu=1}^N p_\nu$.

One constraint qualification to ensure that the GNEP KKT system is a necessary optimality condition is the GNEP-Slater constraint qualification (SCQ).

Definition 1 We say that GNEP-SCQ holds, if for all $\nu = 1, \dots, N$ and for all $x \in K(x)$ it holds:

$$(\nabla_{x^\nu} h_i^\nu(x))_{i=1, p_\nu} \text{ are linearly independent,}$$

and there exists x^0 such that

$$\begin{aligned} g_i^\nu(x^{0,\nu}, x^{-\nu}) &< 0, \forall i = 1, \dots, m_\nu, \\ h_i^\nu(x^{0,\nu}, x^{-\nu}) &= 0, \forall i = 1, \dots, p_\nu. \end{aligned}$$

We refer the interested reader to [4, 5, 14] for more discussion on constraint qualifications for (1).

2 A Pathway to the GNEP

Following [35, 36], we study here an homotopy technique for the KKT formulation of the GNEP. The KKT system (2) can be reformulated as a non-linear equation $H(x, \lambda, s, \mu) = 0$, where $H : \mathbb{R}^n \times \mathbb{R}_+^{2m} \times \mathbb{R}^p \rightarrow \mathbb{R}^{n+2m+p}$ is defined as

$$H(x, \lambda, s, \mu) := \begin{pmatrix} (\nabla_{x^v} \mathcal{L}_v(x, \lambda, \mu))_{v=1, N} \\ g(x) + s \\ h(x) \\ (\lambda_i s_i)_{i=1, m} \end{pmatrix},$$

with $\mathcal{L}_v(x, \lambda, \mu) := \theta_v(x) + g^v(x)^T \lambda^v + h^v(x)^T \mu^v$ denotes the Lagrangian function of the v -th player, and denoting $g(x) := (g^v(x))_{v=1, N}$, $h(x) := (h^v(x))_{v=1, N}$. In the rest of this section, we will consider a parametrized version of this system $H(x, \lambda, s, \mu, t) : \mathbb{R}^n \times \mathbb{R}_+^{2m} \times \mathbb{R}^p \times \mathbb{R}^l \rightarrow \mathbb{R}^{n+2m+p}$ defined as

$$H(x, \lambda, s, \mu, t) := \begin{pmatrix} (\nabla_{x^v} \mathcal{L}_v(x, \lambda, \mu, t))_{v=1, N} \\ g(x, t) + s \\ h(x, t) \\ \Phi(\lambda, s, t) \end{pmatrix}, \quad (3)$$

where $g(x, t)$ and $h(x, t)$ are possible regularizations of $g(x)$ and $h(x)$, and $\Phi : \mathbb{R}^{2m+l} \rightarrow \mathbb{R}^{2m}$ is a smooth homotopy mapping. Examples of such mappings will be given in Sect. 3.

Denote $\mathbf{0}$ (resp. $\mathbf{1}$), the vector of all zeros (resp. ones). We assume that the parametrized system satisfies

$$H(x, \lambda, s, \mu, \mathbf{0}) = H(x, \lambda, s, \mu), \quad (4)$$

and the solution is easy to compute at $t = \mathbf{1}$.

This formulation encompasses some classical smoothing approach for complementarity problems such as smoothing of merit function or primal-dual interior-point methods, as discussed in the next section.

The next result follows straightforwardly from the Implicit Function Theorem.

Theorem 1 *Assume that $\forall (x, \lambda, \mu, s, t) \in \mathbb{R}^n \times \mathbb{R}_+^{2m} \times \mathbb{R}^p \times [0, 1]^l$, $\nabla_{x, \lambda, \mu, s} H(x, \lambda, s, \mu, t)$ has a full rank $(n + 2m + p)$, and the functions θ_v, g^v, h^v are C^3 for all v . Then, $H^{-1} := \{(x, \lambda, \mu, s, t) : H(x, \lambda, s, \mu, t) = 0\}$ is an l -smooth manifold.*

The following lemma gives a more specific condition to ensure the full rank assumption.

Definition 2 We say that GNEP-SCQ(t) holds, if for all $v = 1, \dots, N$ and for all $x \in K(x, t)$ it holds:

$$(\nabla_{x^v} h_i^v(x, t))_{i=1, p_v} \text{ are linearly independent,} \quad (5)$$

and there exists x^0 such that

$$\begin{aligned} g_i^v(x^{0,v}, x^{-v}, t) &< 0, \forall i = 1, \dots, m_v, \\ h_i^v(x^{0,v}, x^{-v}, t) &= 0, \forall i = 1, \dots, p_v. \end{aligned} \quad (6)$$

At $t = \mathbf{0}$, we recover the classical GNEP-SCQ given in previous section.

From now on, assume that the set of feasible points of (1), denoted $X := \{x : x \in K(x)\} \subseteq \mathbb{R}^n$, is a non-empty compact convex set.

Lemma 1 *Assume that for all $x \in X$, $\theta_v(\cdot, x^{-v})$ and $g^v(\cdot, x^{-v})$ is convex, and $h^v(\cdot, x^{-v})$ is affine. Besides, assume that, for any $t \in [0, 1]^l$, GNEP-SCQ(t) holds true. Then, H^{-1} is contained in a compact set.*

Proof Consider each parametric optimization problem composing the GNEP KKT system separately. It is classical from non-linear optimization that the SCQ is equivalent to the Mangasarian Fromovitz constraint qualification (MFCQ) in the convex case. Then, MFCQ is equivalent to having the set of Lagrange multipliers non-empty and bounded, according to [23].

Since GNEP-SCQ consists of applying the classical SCQ to each parametric optimization problem, the previous reasoning gives that the set of Lagrange multiplier is non-empty and bounded. \square

We assume that there exists a compact convex set $D \subseteq X \times \mathbb{R}_+^{2m} \times \mathbb{R}^p$ such that

$$H(x, \lambda, s, \mu, t) = 0 \implies (x, \lambda, \mu, s) \in \text{int}(D) \text{ or } t = 0.$$

In other words, H is boundary-free for $t > 0$. Assuming GNEP-SCQ(t) holds for any $t \in [0, 1]^l$, and X is a compact convex set, then this assumption is not restrictive according to previous lemma.

Theorem 2 *Suppose $H(x, \lambda, s, \mu, t)$ is a homotopy system of $H(x, \lambda, s, \mu)$ such that:*

- (i) *the functions θ_v, g^v, h^v are C^3 , and for all $x \in X$, $\theta_v(\cdot, x^{-v})$ and $g^v(\cdot, x^{-v})$ is convex, and $h^v(\cdot, x^{-v})$ is affine;*
- (ii) *for any $(x, \lambda, \mu, s, t) \in D \times [0, 1]^l$, the jacobian matrix $\nabla_{x,\lambda,\mu,s} H(x, \lambda, s, \mu, t)$ has a full rank $(n + 2m + p)$;*
- (iii) *for all $t \in [0, 1]^l$, GNEP-SCQ(t) holds true.*
- (iv) *At $t = \mathbf{1}$, the system $H(x, \lambda, s, \mu, \mathbf{1})$ has a unique solution.*

The path follows leads from the unique starting point $(x^0, \lambda^0, \mu^0, s^0)$ to a solution of the GNEP.

Proof Theorem 1 guarantees that, starting from the initial point $(x^0, \lambda^0, \mu^0, s^0)$, there exists a path of solutions of $H = 0$. This path cannot diverge by (iii) and Lemma 1. Nor can it go back to the initial point by (iv) and by (ii) it cannot loop. Thus, the path reach $t = \mathbf{0}$, which is a solution of the KKT GNEP system, and is therefore a generalized Nash equilibrium by (i). \square

We refer the reader to [13, Sect. 3] for a discussion on assumptions guaranteeing the full rank of the GNEP KKT system.

Remark 1 In the case where $K_v(x^{-v})$ does not depend on x^{-v} (classical Nash games), one way to ensure the uniqueness of the initial point is to add a regularization term in the objective function of each player in the following way:

$$(1 - t)\theta_v(x^v, x^{-v}) + \frac{t}{2}\|x^v - x^{0,v}\|^2.$$

Further research will focus on weakening some of the assumptions of this theorem. In particular, assumption (ii) implies that the equilibrium is locally unique, which might be restrictive. Additionally, the question of using a weaker assumption than GNEP-SCQ would also be of great interest for the applicability of the homotopy. Another perspective involves the generalization of this scheme to weaker optimality conditions than the GNEP KKT, such as Fritz John system [9] inspired by a recent adaptation of Newton method’s variants [20].

3 Examples of Pathways

In the previous section, we introduced a parametric mapping (3) that can be used in a path-following approach to tackle the GNEP. Throughout this section, we consider one parameter $t \in [0, 1]$ so that $l = 1$.

Given an initial point $x^0 \in X$, we denote the regularized Lagrangian function of the v -th player by

$$\mathcal{L}_v(x, \lambda, \mu, t) := (1 - t)\theta_v(x^v, x^{-v}) + \frac{t}{2}\|x^v - x^{0,v}\|^2 + g^v(x^*)^T \lambda + h^v(x^*)^T \mu.$$

For a smoothing function $\Phi : \mathbb{R}^{2m+1} \rightarrow \mathbb{R}^m$, consider the following specialization of (3):

$$H(x, \lambda, s, \mu, t) := \begin{pmatrix} (\nabla_{x^v} \mathcal{L}_v(x, \lambda, \mu, t))_{v=1,N} \\ g(x) + s \\ h(x) \\ \Phi(\lambda, s, t) \end{pmatrix}. \tag{7}$$

We now discuss two examples of parametric systems by specializing the function Φ . Note that assuming for all $t > 0$ that $\lambda_i s_i = 0 \implies \Phi_i(\lambda, s, t) < 0$, and starting the homotopy ($t = 1$) at $(x^0, \lambda^0, \mu^0, s^0)$ with $\lambda^0 > 0, s^0 > 0$, the continuity of the pathway guarantees that λ and s remains positive at any $t > 0$ and possibly vanishes only at $t = 0$.

3.1 Smoothing of C-Functions

We call Φ a C-function, [16], if it satisfies the following property:

$$\Phi(a, b) = 0 \iff a, b \geq 0, (a_i b_i)_{i=1}^m = 0.$$

This type of function is particularly useful in the context of complementarity problems. Classical C-functions are the min function, $\Phi_{\min}(a, b) = \min(a, b)$, and the Fischer-Burmeister (FB) function, $\Phi_{FB}(a, b) := a + b - \sqrt{a^2 + b^2}$. However, these functions are non-smooth in general and one can use a smoothing techniques to satisfy the assumptions of Theorem 2. For instance, a smoothing of the FB-function is

$$\Phi_{FB}(a, b, t) := a + b - \sqrt{a^2 + b^2 + t^2}.$$

We refer the reader to [7] or [16] and references therein. Noting that previous approach does not exploit the fact that the non-negativity constraints are ensured by the homotopy, we can also consider merit function that only regularize the product $ab = 0$, see for instance [1].

3.2 Interior-Point Method

Following the classical interior-point method in numerical optimization, consider the following

$$\Phi_{IPM}(\lambda, s, t) := (\lambda_i s_i)_{i=1}^m - t.$$

This formulation allows to exploit the very efficient algorithms design for interior-point methods [34]. Moreover, we can also benefit from the good theoretical (polynomial-time) complexity of the interior-point method in some specific cases, see recently [24].

4 How to Follow the Path

Based on the homotopy introduced in the previous sections, we can derive numerical approaches that are following the path. In other words, the method consists of discretizing the interval $[0, 1]^l$ and at each step solving the parametrized problem. Considering a uniform step-size, a generic algorithm is given in Algorithm 1.

It might not be efficient to solve all the subproblems with high precision since we only really care about the final solution. One classical approach taking into account this information is the so-called predictor-corrector method [2]. The idea is to first use a predictor step, for instance, one iteration of a Newton method. Then, run

```

starting vector  $z^0 := (x, \lambda, s, \mu)^0 \in \text{int}(D)$  such that  $H(z^0, \mathbf{1}) = 0$ ;
 $N$  an integer;
 $\epsilon > 0$ ;
Begin ;
Set  $k := 1, \delta := \frac{N-1}{N}, \Delta\delta = \frac{1}{N}$ ;
for  $k = 1, \dots, N$  do
    Compute  $z^k$  iterative solution  $\|H(z, \delta\mathbf{1})\| \leq \epsilon$  using  $z^{k-1}$  as initial point;
     $\delta := \delta - \Delta\delta$ ;
end for
return  $x^k$ 

```

Algorithm 1: Generic path-following algorithm

a corrector step (possibly more than one) to get closer to the path. We left more sophisticated discussion on implementation such as an effective step-size adaptation, an efficient incorporation of higher-order predictors, and an efficient implementation of the corrector step specific to the GNEP KKT system for future works.

Acknowledgements This work was supported by an NSERC Discovery Accelerator Supplement, grant number 401285 of the second author. The authors would like to thank anonymous referees for their helpful remarks and comments.

References

1. Abdallah, L., Haddou, M., Migot, T.: A sub-additive DC approach to the complementarity problem. *Comput. Optim. Appl.* **73**(2), 509–534 (2019)
2. Allgower, E.L., Georg, K.: *Numerical Continuation Methods: An Introduction*, vol. 13. Springer Science & Business Media (2012)
3. Arrow, K.J., Debreu, G.: Existence of an equilibrium for a competitive economy. *Econ. J. Econ. Soc.* 265–290 (1954)
4. Aussel, D., Svensson, A.: Towards tractable constraint qualifications for parametric optimisation problems and applications to generalised Nash games. *J. Optim. Theory Appl.* **182**(1), 404–416 (2019)
5. Bueno, L.F., Haeser, G., Rojas, F.N.: Optimality conditions and constraint qualifications for generalized Nash equilibrium problems and their practical implications. *SIAM J. Optim.* **29**(1):31–54 (2019)
6. Cavazzuti, E., Pappalardo, M., Passacantando, M.: Nash equilibria, variational inequalities, and dynamical systems. *J. Optim. Theory Appl.* **114**(3), 491–506 (2002)
7. Chen, C., Mangasarian, O.L.: A class of smoothing functions for nonlinear and mixed complementarity problems. *Comput. Optim. Appl.* **5**(2), 97–138 (1996)
8. Cojocaru, M.-G., Wild, E., Small, A.: On describing the solution sets of generalized Nash games with shared constraints. *Optim. Eng.* **19**(4), 845–870 (2018)
9. Dorsch, D., Jongen, H.T., Shikhman, V.: On structure and computation of generalized Nash equilibria. *SIAM J. Optim.* **23**(1), 452–474 (2013)
10. Dreves, A.: Improved error bound and a hybrid method for generalized Nash equilibrium problems. *Comput. Optim. Appl.* **65**(2), 431–448 (2016)
11. Dreves, A.: How to select a solution in generalized Nash equilibrium problems. *J. Optim. Theory Appl.* **178**(3), 973–997 (2018)

12. Dreves, A., Facchinei, F., Fischer, A., Herrich, M.: A new error bound result for generalized Nash equilibrium problems and its algorithmic application. *Comput. Optim. Appl.* **59**(1), 63–84 (2014)
13. Dreves, A., Facchinei, F., Kanzow, C., Sagratella, S.: On the solution of the KKT conditions of generalized Nash equilibrium problems. *SIAM J. Optim.* **21**(3), 1082–1108 (2011)
14. Facchinei, F., Kanzow, C.: Generalized Nash equilibrium problems. *Ann. Oper. Res.* **175**(1), 177–211 (2010)
15. Facchinei, F., Pang, J.-S.: Exact Penalty Functions for Generalized Nash Problems, pp. 115–126. Springer US, Boston, MA (2006)
16. Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Science & Business Media (2007)
17. Facchinei, F., Pang, J.-S.: Nash equilibria: the variational approach. In: *Convex Optimization in Signal Processing and Communications*, pp. 443–493. Cambridge University Press (2010)
18. Facchinei, F., Piccialli, V., Sciandrone, M.: Decomposition algorithms for generalized potential games. *Comput. Optim. Appl.* **50**(2), 237–262 (2011)
19. Fan, X., Jiang, L., Li, M.: Homotopy method for solving generalized Nash equilibrium problem with equality and inequality constraints. *J. Ind. Manag. Optim.* **15**(4), 1795–1807 (2019)
20. Fischer, A., Herrich, M.: Newton-type methods for Fritz John systems of generalized Nash equilibrium problems. *Pure Appl. Funct. Anal.* **3**(4), 587–602 (2018)
21. Fischer, A., Herrich, M., Izmailov, A.F., Solodov, M.V.: A globally convergent LP-Newton method. *SIAM J. Optim.* **26**(4), 2012–2033 (2016)
22. Fischer, A., Herrich, M., Schönefeld, K.: Generalized Nash equilibrium problems-recent advances and challenges. *Pesqui. Oper.* **34**(3), 521–558 (2014)
23. Gauvin, J.: A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Math. Program.* **12**(1), 136–138 (1977)
24. Haddou, M., Migot, T., Omer, J.: A generalized direction in interior point method for monotone linear complementarity problems. *Optim. Lett.* **13**(1), 35–53 (2019)
25. Izmailov, A.F., Solodov, M.V.: On error bounds and Newton-type methods for generalized Nash equilibrium problems. *Comput. Optim. Appl.* **59**(1), 201–218 (2014)
26. Kanzow, C., Steck, D.: Quasi-variational inequalities in Banach spaces: theory and augmented Lagrangian methods. *SIAM J. Optim.* **29**(4), 3174–3200 (2019)
27. Migot, T., Cojocaru, M.-G.: A decomposition method for a class of convex generalized Nash equilibrium problems. *Optim. Eng.* (2020)
28. Migot, T., Cojocaru, M.-G.: Nonsmooth dynamics of generalized Nash games. *J. Nonlinear Var. Anal.* **1**(4), 27–44 (2020)
29. Migot, T., Cojocaru, M.-G.: A parametrized variational inequality approach to track the solution set of a generalized Nash equilibrium problem. *Eur. J. Oper. Res.* **283**(3), 1136–1147 (2020)
30. Nash, J.F.: Equilibrium points in n-person games. *Proc. Natl. Acad. Sci.* **36**(1), 48–49 (1950)
31. Nikaidō, H., Isoda, K.: Note on non-cooperative convex games. *Pac. J. Math.* **5**(Suppl. 1), 807–815 (1955)
32. Pang, J.-S., Scutari, G., Facchinei, F., Wang, C.: Distributed power allocation with rate constraints in gaussian parallel interference channels. *IEEE Trans. Inf. Theory* **54**(8), 3471–3489 (2008)
33. Stein, O., Sudermann-Merx, N.: The cone condition and nonsmoothness in linear generalized Nash games. *J. Optim. Theory Appl.* **170**(2), 687–709 (2016)
34. Terlaky, T.: *Interior Point Methods of Mathematical Programming*, vol. 5. Springer Science & Business Media (2013)
35. Zangwill, W.I., Garcia, C.B.: Equilibrium programming: the path following approach and dynamics. *Math. Progr.* **21**(1), 262–289 (1981)
36. Zangwill, W.I., Garcia, C.B.: *Pathways to Solutions, Fixed Points and Equilibria*. Prentice-Hall, Englewood Cliffs, NJ (1981)

Properties of the Zeros of the Scale-Delay Equation and Its Time-Variant ODE Realization



Erik I. Verriest

Abstract An inverse realization problem is solved for a class of analytic functions: Given a function, find a regular differential polynomial that annihilates it. It is shown that the minimal annihilator has degree $m + 1$, where m is the highest multiplicity of the zeros of x belonging to a class of analytic functions. This generalizes the realization of Bohl functions as solutions to LTI-ODE's. With it, the unit solution of the scale-delay equation (SDE) is approximated as the solution to a second-order time-variant ODE. Some new identities for the exact zeros of the SDE are proven.

Keywords Differential polynomial · Inverse problem · ODE modeling · Scale delay equation · Linear time-variant system

1 Introduction

It is well known that the solution of homogeneous finite order linear time-invariant (LTI) ordinary differential equation consists of a finite sum of (complex) exponentials multiplied by polynomials in t . Such a function is known as a Bohl function [7]. Zeilberger calls these C-finite, as the vector space $C[\mathbf{D}]f = \text{span}\{\mathbf{D}^i f; i \geq 0\}$, where \mathbf{D} denotes the differential operator $\frac{d}{dt}$, is finite dimensional [13]. In the real case, a Bohl function is a finite sum of products of polynomials, real exponentials and sines and cosines.

An inverse problem is associated with the previous: Given any Bohl function,

$$x(t) = \sum_{i=1}^{\mu} p_i(t)e^{\lambda_i t}, \quad \lambda_i \in \mathbb{C}, \quad \text{such that } i \neq j \Rightarrow \lambda_i \neq \lambda_j, \quad p_i \in \mathbb{C}[t], \quad \deg p_i = m_i, \quad (1)$$

find an annihilator for $x(t)$, i.e., a polynomial operator $a(\mathbf{D})$, with $a(s) \in \mathbb{R}[s]$, such that $a(\mathbf{D})x = 0$. (Here, $\mathbf{D} = d/dt$). This inverse problem does not have a unique

E. I. Verriest (✉)

Georgia Institute of Technology, Atlanta GA, 30332-0250, USA
e-mail: erik.verriest@ece.gatech.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_10

answer. If $a(\mathbf{D})$ is a solution, i.e., $a(\mathbf{D})x = 0$, then so is $c(\mathbf{D})a(\mathbf{D})$. In the ring $\mathbb{R}[\mathbf{D}]$ this inverse problem specifies an ideal. However, the *monic* solution of least degree is unique. The ideal, annihilating x , is then generated by

$$a_o(\mathbf{D}) \stackrel{\text{def}}{=} \prod_{i=1}^{\mu} (\mathbf{D} - \lambda_i)^{m_i+1}, \quad \deg a_o = \sum_{i=1}^{\mu} (m_i + 1). \quad (2)$$

In this paper we ask the following questions:

1. Can one do better for Bohl functions if one allows time-varying coefficients?
2. Can one find $a(t, \mathbf{D})$ if x is non-Bohl?
3. What is the minimal order of $a(t, \mathbf{D})$ solving the problem in (2)? We will see that the answers to question 1 and 2 are affirmative, under some conditions. We then proceed with the constructive answer to question 3.

Of course, the ODE should be restricted to have smooth coefficients. Hence we formulate our generalized inverse problem, restricted to real functions, as follows.

Given a smooth, at least n times differentiable function, $x(t)$, find a linear differential operator,

$$a(t, \mathbf{D}) = \mathbf{D}^n + a_1(t)\mathbf{D}^{n-1} + \cdots + a_{n-1}\mathbf{D} + a_n(t), \quad (3)$$

with $a_i(t) \in C(\mathbb{R}, \mathbb{R})$ such that $a(t, \mathbf{D})x \equiv 0$ for $t \in (a, b)$. First, we look in Sect. 2 at some examples to shed light on the problem. In Sect. 3, this problem is completely solved for functions that are analytic functions in a finite interval. Section 4 gives an extension to infinite intervals, requiring some notions of entire functions. The last section looks at the solution of the (scalar) scale-delay equation, which can be seen as the simplest generalization (in Weierstrass's sense) of the exponential function, and some new properties of the zeros of this function are described.

2 Time-Variant Differential Annihilators

Zeilberger introduced the notion of a holonomic function [13]. It is a function that is annihilated by a polynomial in \mathbf{D} with polynomial (in t) coefficients. Define the operation of *multiplication by the independent variable* by \mathbf{Q} . The differential polynomials considered are then elements from the noncommutative Weyl algebra generated by \mathbf{Q} and \mathbf{D} . Thus $x \in \ker(\mathbb{R}[\mathbf{Q}][\mathbf{D}])$ iff x is holonomic. In this work, we shall also allow the coefficients of the differential polynomials to be rational functions, or power series. Let's look at some motivating examples.

First, consider nonvanishing functions. Given $x \in C^1((\alpha, \beta), \mathbb{R})$ such that $x(t) \neq 0$ in (α, β) , then let

$$a(t, \mathbf{D}) = \mathbf{D} - \frac{\dot{x}(t)}{x(t)} \quad (4)$$

which involves the logarithmic derivative. This form is well-defined and solves the problem since $(\mathbf{D} - \frac{\dot{x}}{x})x = 0$.

Next, consider the case of a function x having a single simple zero in an interval. If $x \in C^1((\alpha, \beta), \mathbb{R})$ vanishes at $t_0 \in (\alpha, \beta)$, the first order operator $a(t, \mathbf{D})$ in (4) is not well-defined. Restricted to (α, t_0) , all solutions of (4) are of the form $Ax(t)$, $A \in \mathbb{C}$. Likewise, in (t_0, β) , the solutions are of the form $Bx(t)$, $B \in \mathbb{C}$. It is not necessary to assume $A = B$.

Example 1 To shed light on the behavior when $x(t)$ has a simple zero, consider $x(t) = t - 1$, in an interval (α, β) containing 1. The naive form leads to $a(t, \mathbf{D}) = (\mathbf{D} - \frac{1}{t-1})$, and the behaviors consistent with this are:

$$\left(\mathbf{D} - \frac{1}{t-1}\right)y = 0 \Rightarrow y(t; A, B) = \begin{cases} A(t-1) & \text{if } t \in (0, 1) \\ B(t-1) & \text{if } t \in (1, \infty). \end{cases}$$

Thus, the corresponding Cauchy problem has (weak) solutions (with $H(\cdot)$ the Heaviside unit-step function):

$$A + (B - A)H(t - 1)](t - 1). \tag{5}$$

Two parameters pin down a particular solution, which alludes to a *higher dimensionality*. With the identity $t\delta(t) = 0$, it is readily verified that the functions (5) are indeed all annihilated by the non-monic differential operator $(t - 1)\mathbf{D} - 1$. A non-monic ODE where the highest order coefficient can vanish is known as singular.

Example 2 Consider the function $x(t) = t + t^2H(t)$. It has a single zero at $t = 0$, and is differentiable with derivative $\dot{x}(t) = 1 + 2tH(t)$. Its second derivative is not continuous at zero. Multiplying by t , we obtain the differential polynomial

$$a(t, \mathbf{D}) = t\mathbf{D} - \frac{1 + 2tH(t)}{1 + tH(t)}. \tag{6}$$

The equation $a(t, \mathbf{D})y(t) = 0$ has the general solution $y(t) = At$ for $t < 0$, and $y(t) = Bt(1 + t)$ for $t > 0$, hence is continuous. However, its first derivative has a jump of $B - A$ at $t = 0$, and its second derivative has the singular term $(B - A)\delta(t)$ and a jump $2BH(t)$. Unless $y \equiv 0$, smoothness is at most of first order. The next section explores singular cases in more detail.

3 General Results

The examples in Sect. 2 suggest an idea on how to generate a singular time-variant first order ODE in case the given function x has a single zero in some interval (α, β) . They also showed that it may be prudent to limit the analysis to *analytic* functions. The first main result proven here relates to arbitrary non-identically zero analytic

$x(t)$. By the principle of permanence, this implies that the zeros of x are isolated (no cluster points). Consequently, a real analytic function has finitely many zeros in any bounded interval.

Definition 1 The linear differential polynomial $a(t, \mathbf{D}) = \mathbf{D}^n + a_1(t)\mathbf{D}^{n-1} + \dots + a_{n-1}\mathbf{D} + a_n(t)$ is *regular* in (α, β) if it is monic ($a_0(t) \equiv 1$) and the coefficients $a_i(t)$ are *continuous* in (α, β) .

Definition 2 A smooth function is called *signed* in (α, β) if it is nowhere vanishing in the interval (α, β) . Consequently, it has a fixed sign in (α, β) .

We shall now restrict the given function x to be real analytic. Recall that x is real analytic on (α, β) iff x can be extended to a complex analytic (a.k.a. holomorphic) function on an open set $\mathcal{D} \subset \mathbb{C}$, which contains the real interval (α, β) . The set of real analytic functions on an interval (α, β) is denoted by $C^\omega((\alpha, \beta), \mathbb{R})$.

Theorem 1 If $x \in C^\omega((\alpha, \beta), \mathbb{R})$ has only single zeros, then there exists a regular second order linear differential operator $a(t, \mathbf{D})$ such that $a(t, \mathbf{D})x = 0$.

Proof Let the zeros of x in (α, β) be $\alpha < t_1 < t_2 < \dots < t_n < \beta$, and factor $x(t)$ as

$$x(t) = \underbrace{(t - t_1)(t - t_2) \cdots (t - t_n)}_{=p(t)} x_r(t),$$

where $x_r(t)$ is twice differentiable and $p(t)\mathbf{D} - \frac{p(t)\dot{x}_r(t)}{x_r(t)} - \dot{p}(t)$, annihilates the given function x but is not regular. Operate on the left with the first-order differential polynomial $\mathbf{D} - \eta(t)$, where $\eta \in C((\alpha, \beta), \mathbb{R})$, to get

$$\begin{aligned} & (\mathbf{D} - \eta) \left(p\mathbf{D} - \frac{p\dot{x}_r}{x_r} - \dot{p} \right) \\ &= p\mathbf{D}^2 - p \left(\eta + \frac{\dot{x}_r}{x_r} \right) \mathbf{D} + p \left(\frac{\dot{x}_r}{x_r} \eta - \left(\frac{\dot{x}_r}{x_r} \right)' \right) + \left[\dot{p} \left(\eta - \frac{\dot{x}_r}{x_r} \right) - \ddot{p} \right]. \end{aligned}$$

If the term in $[\cdot]$ were a multiple of p , we may cast out the “ p ” from the above differential operator to obtain a monic one. Thus motivated, we ask for the solvability for η and k in

$$\dot{p} \left(\eta - \frac{\dot{x}_r}{x_r} \right) - \ddot{p} \stackrel{?}{=} kp. \tag{7}$$

Note that by the Gauss-Lucas theorem, the polynomials p and \dot{p} have interlaced roots. This implies that p and \dot{p} are coprime polynomials. By Bezout’s theorem, polynomials q_0 and k_0 exists in $\mathbb{R}[t]$ such that

$$q_0(t)\dot{p}(t) - k_0(t)p(t) = 1. \tag{8}$$

This is a consequence of the Euclidean division algorithm. The Diophantine equation

$$q(t)\dot{p}(t) - k(t)p(t) = \ddot{p}(t) \tag{9}$$

is then solved by the polynomials $q(t) = q_0(t)\ddot{p}(t)$ and $k(t) = k_0(t)\ddot{p}(t)$. Finally, let

$$\eta(t) = q(t) + \frac{\dot{x}_r(t)}{x_r(t)},$$

so that the regular second order differential polynomial,

$$a(t, \mathbf{D}) = \mathbf{D}^2 - \left(q_0\ddot{p} + 2\frac{\dot{x}_r}{x_r} \right) + \left(\left(\frac{\dot{x}_r}{x_r} \right) q_0\ddot{p} + \left(\frac{\dot{x}_r}{x_r} \right)^2 - \left(\frac{\dot{x}_r}{x_r} \right)' + k_0\ddot{p} \right), \tag{10}$$

annihilates $x(t) = p(t)x_r(t)$. □

Theorem 1 explains why in the theory of linear time-variant equations most of the study centers around *second-order* equations, culminating in the Sturm-Liouville problem. Furthermore, single zero-crossings are robust with respect to perturbations. Higher order zeros are not. We present here an alternative solution method, based on the following lemma.

Lemma 1 *Let p be a monic polynomial of degree n with single real roots t_1, \dots, t_n . Then the Bezout equation $q_0(t)\dot{p}(t) - k_0(t)p(t) = 1$ is solved by*

$$q_0(t) = Q_0t^{n-1} + Q_1t^{n-2} + \dots + Q_{n-1}, \tag{11}$$

where the coefficients, Q_i , are obtained by solving the system of equations, parameterized by the real zero set, $\{t_1, \dots, t_n\}$,

$$\begin{bmatrix} t_1^{n-1} & \dots & t_1 & 1 \\ t_1^{n-1} & \dots & t_1 & 1 \\ \vdots & & \vdots & \\ t_1^{n-1} & \dots & t_1 & 1 \end{bmatrix} \begin{bmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_{n-1} \end{bmatrix} = \begin{bmatrix} q_0(t_1) \\ q_0(t_2) \\ \vdots \\ q_0(t_n) \end{bmatrix}. \tag{12}$$

where $q_0(t_i) = \text{Res} \left[\frac{1}{p(t)}, t_i \right]$. The matrix on the left of (12) is a Vandermonde matrix, and is nonsingular since the t_i are disjoint. The polynomial $k_0(t)$ is then identified from the given Bezout identity.

Proof A polynomial, p , with single real roots at t_i , has the property $\frac{\dot{p}(t)}{p(t)} = \sum_i \frac{1}{t-t_i}$. Hence the Bezout equation (8) is equivalent to $q_0(t) \sum_{i=1}^n \frac{1}{t-t_i} - k_0 = \frac{1}{p(t)}$, from which $q_0(t_k) \sum_{i=1}^n \prod_{j \neq i} (t_k - t_j) = 1$. Only one term in this sum is non-zero, so it follows from

$$q_0(t_k) = \frac{1}{\prod_{j \neq k} (t_k - t_j)}$$

that $q_0(t_i) = \lim_{t \rightarrow t_i} \frac{t-t_i}{p(t)}$, i.e., the $q_0(t_i)$ are the residues of the inverse of p at its roots, or equivalently, the numerators in the partial fraction expansion of p^{-1} . The coefficients Q_i are determined by the solution of the Vandermonde system (12). \square

In Examples 1 and 2, the function $x(t)$ is differentiable, and in both cases x has a nonzero derivative where x vanishes. Hence the zero has multiplicity one in both cases. In Example 1, the choice $\eta = 0$ yields the second order regular differential polynomial \mathbf{D}^2 . In example (2) it is not possible to find a second order regular polynomial, since x is not twice differentiable. This is the reason for restricting Theorem 1 to analytic functions. Next we generalize to real analytic x possessing real zeros of higher multiplicity.

Lemma 2 Any $x \in C^\omega((\alpha, \beta), \mathbb{R})$ can be factored as $x = px_b$, where p is a monic polynomial with only real zeros in (α, β) and x_b is differentiable and signed in (α, β) . Then

$$\left(\mathbf{D} - \frac{\dot{x}_b}{x_b} \right) px_b = \dot{p}x_b.$$

Proof Direct verification. \square

The factorization alluded to in Lemma 2 is not unique. The function $x(t) = t(t^2 + 1) \exp(-t)$ factors in $p_1(t) = t$ and $x_{b1}(t) = (t^2 + 1) \exp(-t)$, or $p_2(t) = t(t^2 + 1)$ and $x_{b2}(t) = \exp(-t)$ in the interval $(-1, 1)$. This prompts us to define a canonical factorization:

Definition 3 The factorization of $x \in C^\omega((\alpha, \beta), \mathbb{R})$ as px_b is *canonical* in (α, β) if p is monic and its extension over \mathbb{C} has no roots other than those in the real interval (α, β) . This implies that the cofactor x_b is signed in (α, β) .

By the Gauss-Lucas theorem, all roots of \dot{p} lie on the real axis, and $\deg \dot{p} = \deg p - 1$. In addition, if p has a root of multiplicity m at $t = t_0 > 1$, then \dot{p} has a root at t_0 of multiplicity $m - 1$. This leads to:

Theorem 2 Let $x = px_b$ be a canonical factorization of x . If the highest multiplicity of a root is m , then a regular differential polynomial annihilating x is of the form

$$a(t, \mathbf{D}) = (\mathbf{D}^2 + a_1\mathbf{D} + a_2) \left(\mathbf{D} - \frac{\dot{x}_b}{x_b} \right)^{m-1}. \tag{13}$$

Proof Repeated use of Lemma 2 gives $\left(\mathbf{D} - \frac{\dot{x}_b}{x_b} \right)^{m-1} px_b = qx_b$, where $q = p^{(m-1)}$ has only roots on the real axis with multiplicity one. By Theorem 1, smooth functions a_1 and a_2 exist such that qx_b is annihilated by $a(t, \mathbf{D})$. \square

Corollary 1 If the highest multiplicity of a zero of $x(t) \in C^\omega((\alpha, \beta), \mathbb{R})$ is m , then x is annihilated by a regular differential polynomial of degree $m + 1$.

4 Extensions: From (α, β) to \mathbb{R}

If x has a finite number of real zeros, the previous extends directly for $(\alpha, \beta) \rightarrow \mathbb{R}$. But when x has infinitely many zeros, the factor $p(t)$ does not make sense. Still assuming that x is analytic, the zeros cannot cluster and consequently, if $\{t_i\}$ is the sequence of zeros, $|t_n| \rightarrow \infty$. The infinite product $\prod_{n=1}^{\infty} \left(1 - \frac{t}{t_n}\right)$ converges if $\sum_{n=1}^{\infty} \frac{1}{|t_n|}$ converges. Here we invoke Weirstrass’s factorization theorem:

If x is an entire function with zeros, z_i , possibly repeated and for some integer sequence $\{p_n\}$ it holds that $\sum_{n=1}^{\infty} \left(\frac{t}{|z_n|}\right)^{1+p_n}$ converges $\forall t \in \mathbb{R}$, then $x(z) = g(z) \prod_{n=1}^{\infty} E_{p_i} \left(\frac{z}{z_i}\right)$, where

$$E_0(z) = (1 - z), \quad E_p(z) = (1 - z) \exp\left(z + \frac{z^2}{2} + \dots + \frac{z^p}{p}\right), \quad p = 1, 2, \dots$$

are Weierstrass’s elementary factors [6], and g is an entire function without real zeros. The elementary factors are close to 1 if $|z| < 1$, and p is large, although $E_p(1) = 0$.

Consider the class of real entire functions (Real means $f : \mathbb{R} \rightarrow \mathbb{R}$) of the form

$$f(z) = C e^{-az^2 + bz} z^m \prod_{k=1}^{\infty} \left(1 - \frac{z}{z_k}\right) e^{z/z_k},$$

with $a \geq 0, b \in \mathbb{R}, C \in \mathbb{R}, \sum_k \frac{1}{|z_k|^2} < \infty, z_k \in \mathbb{C} \setminus \{0\}, |\text{Im } z_n| < \infty$. Functions in this class have only real zeros. Examples are $\sin(z), \cos(z), \exp(z), \exp(-z)$, and $\exp(-z^2)$. This class is known as the *Laguerre-Pólya class*, denoted LP. We can now make the proper extension for real analytic x .

Theorem 3 *If x , a real entire function, has a factorization $x = x_e \Pi$, with $\Pi \in \text{LP}$ and x_e has no real zeros, then $x(t)$ satisfies a regular second order ODE in all of \mathbb{R} if all zeros of Π have multiplicity one.*

Proof Let $Z(x)$ denote the zero-set for x , and let $\mathcal{T} = Z(x) \cap \mathbb{R}$ be the set of real zeros and $Z_{\mathbb{C}}$ its complement in $Z(x)$. By the factorization theorem

$$x(t) = \underbrace{\prod_{t_n \in \mathcal{T}} E_{p_n} \left(\frac{t}{t_n}\right)}_{=\Pi(t)} \underbrace{\prod_{z_m \in Z_{\mathbb{C}}} E_{p_m} \left(\frac{t}{z_m}\right)}_{=x_e(t)} y(t) \tag{14}$$

where y is entire without zeros. Consequently, x_e , having no zeros in \mathbb{R} is signed. As in Theorem 1, we try to find a smooth function η such that $(\mathbf{D} - \eta) \left(\Pi \mathbf{D} - \dot{\Pi} - \Pi \frac{\dot{x}_e}{x_e}\right)$ is regular. This subproblem requires the solution of the Bezout equation

$$\left(\eta - \frac{\dot{x}_e}{x_e}\right) \dot{\Pi} = -k\Pi = \ddot{\Pi}.$$

The ring of entire functions is a Bezout domain, but not a PID. However, every *finitely* generated ideal is principal. Thus a solution exists if Π and $\dot{\Pi}$ are relatively prime. By Laguerre’s theorem on separation of zeros, this is the case if $\Pi(t)$ belongs to the Laguerre-Polya class [1]. The proof then follows as in Theorem 1. \square

Remark 1 There is no division algorithm to compute η and k for entire functions. However, if $x \in \text{LP}$ has only single zeros, then \dot{x} is well-defined, and does not vanish at the zeros of x . A solution (a, b) to the Bezout equation $ax + b\dot{x} = 1$ is given by $a = \frac{x}{x^2 + \dot{x}^2}$, and $b = -\frac{\dot{x}}{x^2 + \dot{x}^2}$.

Example 3 Consider the Bessel function, $x(t) = J_0(t)$, and recall that $\dot{J}_0(t) = -J_1(t)$ and by Bessel’s ODE: $\ddot{J}_0(t) = -J_0(t) + \frac{J_1(t)}{t}$. Applying Theorem 3 yields

$$(Dl - \eta(t))(J_0(t)\mathbf{D} + J_1(t)) = J_0(t)\mathbf{D}^2 - \eta(t)J_0(t)\mathbf{D} + \dot{J}_1(t) - \eta(t)J_1(t).$$

Solve the Bezout equation, $\eta(t)J_1(t) + k(t)J_0(t) = \dot{J}_1(t)$ to get

$$\eta(t) = \frac{J_1(t)\dot{J}_1(t)}{J_0^2(t) + J_1^2(t)} + p(t)J_0(t), \quad k(t) = \frac{J_0(t)\dot{J}_1(t)}{J_0^2(t) + J_1^2(t)} - p(t)J_0(t).$$

The reason why the usual Bessel ODE for J_0 does not appear stems from the singularity of the second order Bessel ODE: The monic Bessel ODE has a $1/t$ -coefficient.

Example 4 Let now $x(t) = \sin t^2$. This has a double root at $t = 0$, all other roots $(\pm k\sqrt{\pi})$ for $k = 1, 2, \dots$ being simple. In an interval not containing 0, we expect a second order ODE. Indeed, we can find $\mathbf{D}^2 - \frac{1}{t}\mathbf{D} + 4t^2$. In order to find a differential polynomial that is valid in all of \mathbb{R} , consider again $(\mathbf{D} - \eta(t))(t \cos t \mathbf{D}^2 - \mathbf{D} + 4t^3)$. Letting $\eta(t) = t\eta_0(t)$, we get the monic differential polynomials $\mathbf{D}^3 - t^2\eta_0(t)\mathbf{D}^2 + 4(\eta_0(t) + t^2)\mathbf{D} - 12t$, with its simplest form for $\eta_0(t) \equiv 0$, given by $\mathbf{D}^3 + 4t^2\mathbf{D} + 12t$.

5 Application: Scale-Delay Equation

The scale-delay equation, $\dot{x}(t) = Ax(t) + Bx(\alpha t)$ where $\alpha \in (0, 1)$, corresponds to a delay equation, with delay $\tau(t) = (1 - \alpha)t$, and therefore satisfies the causality condition $\dot{\tau} < 1$ [9], and is infinite dimensional. In the limit cases, $\alpha \in \{0, 1\}$, the system is finite dimensional. This equation has been studied extensively [3–5, 8, 11, 14, 15]. We first present some known results to set the stage.

Valeev [8] showed that the scalar functional differential equation (FDA)

$$\dot{y}(t) = \mu y(t) + \beta y(\alpha t), \quad y(0) = 1,$$

has a solution given by the series expansion $y(t) = 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} \prod_{i=0}^{k-1} (\mu + \beta\alpha^i)$. The ratio test shows that this series has an infinite radius of convergence, hence is an entire function. For $\mu \neq 0$, the solution diverges if $|\beta| > |\mu|$ and converges if $|\beta| < |\mu|$. A particularly interesting case is $\mu = 0, \beta = 1$: The unit solution ($y(0) = 1$) (a.k.a. the *deformed exponential*) is an entire function of the LP-class

$$y(t) = \sum_{k=0}^{\infty} \frac{\alpha^{k(k-1)/2}}{k!} t^k \stackrel{\text{def}}{=} E_{\alpha}(t). \tag{15}$$

which satisfies $0 < y(t) < \exp(t^{\epsilon})$ for all $\epsilon > 0$, and $y(t) \geq t^{\frac{\ln \ln t}{2 \ln \alpha} + o(\ln t)}$. It follows that for $\mu = 0$ and arbitrary β , the unit solution is the time-scaled version $E_{\alpha}(\beta t)$. Perhaps surprisingly, the solution $E_{\alpha}(-t)$ for $\beta = -1$ oscillates and diverges. Its zeros are asymptotically given by $t_k = \frac{k}{\alpha^{k-1}}(1 + \psi(\alpha)k^{-2} + o(k^{-2}))$, where $\psi(\alpha)$ is the generating function of the sum-of-divisors function $\sigma(k)$ [12, 15]. Since the order, $\inf_{r>0}\{E_{\alpha}(-z) \sim O(\exp|z|^r)\}$, of $E_{\alpha}(z)$ is zero, Hadamard's factorization theorem yields the simple form in terms of the roots $\{t_k > 0\}$ of $E_{\alpha}(-t)$:

$$E_{\alpha}(-t) = \prod_{k=1}^{\infty} \left(1 - \frac{t}{t_n}\right). \tag{16}$$

Consider the series of inverse powers of the roots

$$S_n = \sum_{k=1}^{\infty} \frac{1}{t_k^n}, \quad n = 1, 2, \dots$$

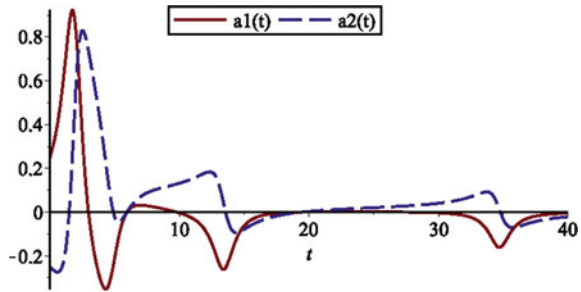
Let also c_n denote the coefficient of t^n in $E_{\alpha}(t)$. Using the extensions of Newton's identities for Weierstrass products, [2] one finds

$$S_n - c_1 S_{n-1} + c_2 S_{n-2} - \dots + (-1)^{n-1} c_{n-1} S_1 + (-1)^n n c_n S_0 = 0, \quad S_0 = 1. \tag{17}$$

Theorem 4 *The sums of inverse powers of the zeros, S_n , can be recursively computed, and give in particular the sequential relations [5]*

$$\begin{aligned} S_1 &= 1 \\ S_2 &= 1 - \alpha \\ S_3 &= \frac{1}{2}(1 - \alpha)^2(2 + \alpha) \\ S_4 &= \frac{1}{6}(1 - \alpha)^3(6 + 6\alpha + 3\alpha^2 + \alpha^3). \end{aligned}$$

Fig. 1 Coefficients a_1 and a_2 of the LTV-ODE equivalent to the scale delay equation ($\alpha = 0.5$)



In addition, the zeros satisfy

$$\sum_{k=1}^{\infty} \frac{1}{\alpha t_k - t_\ell} = 0, \quad \ell = 1, 2, \dots; \quad \frac{1}{t_k} \prod_{\ell \neq k} \frac{t_\ell - t_k}{t_\ell - \alpha t_k} = 1 - \alpha. \quad (18)$$

Proof Substitute Hadamard’s expansion in the FDE, and evaluate at $t = \frac{t_\ell}{\alpha}$ and t_k respectively. □

Since $\dot{x}(t) = -x(\alpha t)$, for $0 < \alpha < 1$ has only positive real zeros all with multiplicity one, by Theorem 3 x must obey a second order linear time-variant ODE.

$$\ddot{x}(t) + a_1(t)\dot{x}(t) + a_2(t)x(t) = 0.$$

Numerical solution of the FDE leads to the approximation of a_1 and a_2 in Fig. 1.

Acknowledgements The support by the NSF grant CPS-1544857 is gratefully acknowledged.

References

1. Boas, R.P.: Entire Functions. Academic Press (1954)
2. Breuer, F.: Identity for Weierstrass product. Am. Math. Monthly **119**(9), 796–799 (2012)
3. Derfel, G., Grabner, P.J., Tichy, R.F.: On the asymptotic behaviour of the zeros of the solutions of a functional-differential equation with rescaling. In: Alpay, D., Kirstein, B. (Eds.) Indefinite Inner Product Spaces, Schur Analysis, and Differential Equations: A Volume Dedicated to Heinz Langer, pp. 281–295. Springer-Verlag (2018)
4. Kato, T., McLeod, J.B.: The functional equation $y'(x) = ay(\lambda x) + by(x)$. Bull. AMS **17**(6), 891–937 (1971)
5. Liu, Y.: On some conjectures by Morris et al. about zeros of an entire function. J. Math. Anal. Appl. **226**, 1–5 (1998)
6. Rudin, W.: Real and Complex Analysis, 2nd ed. McGraw Hill (1974)
7. Trentelman, H., Stoorvogel, A.: Control Theory for Linear Systems. Springer-Verlag (2002)
8. Valeev, K.G.: Linear differential equations with linear time-delay. Siberian Math. J. **5**(2), 290–309 (in Russian)

9. Verriest, E.I.: Inconsistencies in systems with time varying delays and their resolution. *IMA J. Math. Control Inform.* (Special issue: “Time-Delay Systems and Their Applications”) **28**, 147–162 (2011)
10. Verriest, E.I.: Algebraic theory for time variant linear systems: modes, minimality and reachability and observability of interconnected systems. In: *Proceedings of the 32nd IEEE Conference on Decision and Control*, San Antonio, TX, pp. 1349–1354, December 1993
11. Verriest, E.I.: Stability of systems with varying scale delays. In: *Proceedings of the IFAC Symposium on System Structure and Control*, Prague, Czech Republic, August 29–31, 2001
12. Wang, L., Zhang, C.: Zeros of the deformed exponential function. [arXiv:1709.04357v1](https://arxiv.org/abs/1709.04357v1), 13 Sep 2017
13. Zeilberger, D.: A holonomic systems approach to special function identities. *J. Comput. Appl. Math.* **32**, 321–368 (1990)
14. Zhabko, A.P., Laktionov, A.A., Zubov, V.I.: Robust stability of differential-difference systems with linear time-delay, pp. 97–101. Budapest, Hungary, *IFAC Robust Control Design* (1997)
15. Zhang, C.: An asymptotic formula for the zeros of the deformed exponential function. *J. Math. Anal. Appl.* **441**(2), 565–573 (2016)

Advances in Statistical Modelling and Data Analysis

Covering Large Complex Networks by Cliques—A Sparse Matrix Approach



W. M. Abdullah, S. Hossain, and M. A. Khan

Abstract The Edge Clique Cover (ECC) problem is concerned with covering edges of a graph with the minimum number of cliques, which is an NP-hard problem. This problem has many real-life applications, such as, in computational biology, food science, efficient representation of pairwise information, and so on. In this work we propose using a compact representation of network data based on sparse matrix data structures. Building upon an existing ECC heuristic due to Kellerman we proffer adding vertices during the clique-growing step of the algorithm in judiciously chosen degree-based orders. On a set of standard benchmark instances our ordered approach produced smaller sized clique cover compared to unordered processing.

Keywords Adjacency matrix · Clique cover · Intersection matrix · Vertex ordering · Sparse graph

1 Introduction

Identification of and computation with dense or otherwise highly connected subgraphs are two of the kernel operations arising in areas as diverse as sparse matrix determination and complex network analysis [1, 6, 9]. Identification of special interest groups or characterization of information propagation are examples of frequently performed operations in social networks [8]. Efficient representation of network data is critical to addressing algorithmic challenges in the analysis of massive data sets using graph theoretic abstractions. In this paper, we propose sparse matrix data

W. M. Abdullah (✉) · S. Hossain
University of Lethbridge, Alberta, Canada
e-mail: w.abdullah@uleth.ca

S. Hossain
e-mail: shahadat.hossain@uleth.ca

M. A. Khan
Inbridge, Alberta, Canada
e-mail: muhammad@inbridgeinc.com

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_11

structures to enable compact representation of graph data and use an existing sparse matrix framework [5] to design efficient algorithms for the ECC problem.

Let $G = (V, E)$ be an undirected connected graph with $|V| = n$ vertices and $|E| = m$ edges. A clique is a subset of vertices such that every pair of distinct vertices are connected by an edge in the induced (by the subset of vertices) subgraph. An edge clique cover of size k in graph G is a decomposition of set E into k subsets C_1, C_2, \dots, C_k such that $C_i, i = 1, 2, \dots, k$ induces a clique in G and each edge $\{u, v\} \in E$ is included in some C_i . A trivial clique cover can be specified by the set of edges E with each edge being a clique. The problem of finding a clique cover with minimum number of cliques (and many variants thereof) is known to be NP-hard [7].

In the literature, the ECC problem and its variants have been extensively investigated from theoretical perspectives and have found applications in disparate areas. In [3], the authors describe a branch-and-bound approach to determine sparse Jacobian matrices. Given the sparsity pattern of the Jacobian, the problem is to find a partition of the columns into structurally orthogonal column groups of smallest cardinality. Blanchette et al. [15] study the protein complex identification problem from computational biology, where the problem is to identify overlapping protein complexes in protein-protein interaction networks. When modelled as a graph problem, the goal is to decompose the network into a smallest collection of cliques. Several polynomial time algorithms have been proposed in the paper for graphs with bounded tree-width. In sensory science, a seemingly unrelated application area, a frequently occurring task is concerned with the concise representation of pairwise interaction of products with many attributes [14, 16]. This pairwise information can be given in a tabular form called ‘‘compact letter display’’. The challenge is to minimize redundant information. It has been shown that this problem can be posed as a variant of the ECC problem [16].

Many heuristics have been proposed in the literature to approximately solve ECC problem while there are only few exact methods which are usually limited to solving small instance sizes. A recent approach is described by Gramm et al. in [10], where they introduce and analyze data reduction techniques to shrink the instance size without sacrificing the optimal solution. The main idea is that with small enough instance sizes, exact algorithms may become feasible.

In this paper we propose a compact representation of network data based on sparse matrix data structures [3] and provide an improved algorithm based on an existing heuristic for finding clique cover. Our approach is based on the simple but critical observation that for a sparse matrix $A \in \mathbb{R}^{m \times n}$, the column intersection graph of A is isomorphic to the adjacency graph of $A^T A$, and that the row intersection graph of A is isomorphic to the adjacency graph of AA^T [5]. Consequently, the subset of columns corresponding to nonzero entries in row i induces a clique in the adjacency graph of $A^T A$, and the subset of rows corresponding to nonzero entries in column j induces a clique in the adjacency graph of AA^T . Note that, matrices $A^T A$ and AA^T are most likely dense even if matrix A is sparse. In this work, we exploit the connection between sparse matrices and graphs in the reverse direction. We show that given a graph (or network), we can define a sparse matrix, *intersection matrix*, such that graph algorithms of interest can be expressed in terms of the associated

intersection matrix. This structural reduction enables us to use existing sparse matrix computational framework to solve graph problems [5]. This duality between graphs and sparse matrices has also been exploited where the graph algorithms are expressed in the language of sparse linear algebra [1, 4]. However, they use adjacency matrix representation which is different from our intersection matrix representation.

We organize the rest of the paper in the following way. In Sect. 2, we present our main theoretical result that allows us to pose the ECC problem as a matrix determination problem. This is followed by a brief description of the clique-cover heuristic of [11]. Next, we describe algorithms for preprocessing the vertices according to their degree in the graph. Results from numerical experiments on a standard collection of test instances are provided in Sect. 3. Finally, the paper is concluded in Sect. 4.

2 Compact Representation and Edge Clique Cover

Classical data structures adjacency matrix (full matrix storage) and adjacency list for representing graphs are inadequate for efficient computer implementation of many important graph operations. Adjacency matrix is costly for sparse graphs and typical adjacency list implementations employ pointers where indirect access leads to poor cache utilization. In a typical adjacency list implementation of undirected graphs, each edge is represented twice. An alternative adjacency list representation of undirected labelled graph avoids this redundancy by storing each edge only once, where the edges incident on each vertex are stored in sorted order of vertex labels [17]. The intersection matrix representation below enables efficient representation of pairwise information where the edges are implicit. Moreover, it allows us to utilize computational framework DSJM to implement the ECC algorithms.

2.1 Intersection Matrix

We require some preliminary definitions. The *adjacency graph* associated with a matrix $J \in \mathbb{R}^{n \times n}$ is a graph $G = (V, E)$ in which for each column or row k of J there is a vertex $v_k \in V$ and $J(i, j) \neq 0$ implies $\{v_i, v_j\} \in E$. The *column intersection graph* associated with matrix $J \in \mathbb{R}^{m \times n}$ is a graph $G = (V, E)$ in which for each column k of J there is a vertex $v_k \in V$ and $\{v_i, v_j\} \in E$ whenever there is a row l for which $J(l, i) \neq 0$ and $J(l, j) \neq 0$.

Let $G = (V, E)$ be an undirected and connected graph without self-loops or multiple edges between a pair of vertices. The adjacency matrix $A(G) \equiv A \in \{0, 1\}^{|V| \times |V|}$ associated with graph G is defined as,

$$A(i, j) = \begin{cases} 1 & \text{if } \{v_i, v_j\} \text{ where } i \neq j \text{ is in } E \\ 0 & \text{otherwise} \end{cases}$$

Unlike the adjacency matrix which is unique (up to a fixed labeling of the vertices) for graph G , there can be more than one (*column*) *intersection matrix* associated with graph G . We exploit this flexibility to store a graph in a structured and space-efficient form using an intersection matrix. Let the edges in E be labelled $e_1, \dots, e_{|E|}$. An intersection matrix associated with graph $G = (V, E)$ where $|V| = n$ and $|E| = m$, is a matrix $C \in \{0, 1\}^{m \times n}$ where for edge $e_k = \{v_i, v_j\}$, $k = 1, \dots, m$ we have $C(k, i) = C(k, j) = 1$, and all other entries of matrix C are zero.

Let $C \in \{0, 1\}^{m \times n}$ be the intersection matrix as defined above associated with a graph $G = (V, E)$. Consider the product $B = C^\top C$.

Theorem 1 *The adjacency graph of matrix B is isomorphic to graph G .*

Proof Consider an arbitrary edge $e_k = \{v_i, v_j\}$ of graph G . By construction, row k of the intersection matrix C has $C(k, i) = C(k, j) = 1$ and $C(k, l) = 0$ for $l \notin \{i, j\}$. Since there are no multiple edges in G , there is one and only one such row k corresponding to edge e_k . Element $B(i, j)$ is the inner product of column vectors i and j of matrix C . The inner product is 1 if and only if $C(k, i) = C(k, j) = 1$. Thus, e_k is in E if and only if $B(i, j) = 1$ implying that it is an edge connecting vertices v_i and v_j of the adjacency graph of matrix B . This proves the theorem. \square

Theorem 1 establishes the desired connection between a graph and its sparse matrix representation. For a vertex $v \in V$ we define by $N_v = \{w \in V \mid \{v, w\} \in E\}$ the set of its neighbors. The *degree* of a vertex v , denoted $d(v)$, is the cardinality of set N_v . The following result follows directly from Theorem 1.

Corollary 1 *The diagonal entry $B(i, i)$ where $B = C^\top C$ and C is the intersection matrix of graph G , is the degree $d(v_i)$ of vertex $v_i \in V$, $i = 1, \dots, n$ of graph $G = (V, E)$.*

Intersection matrix C defined above represents an edge clique cover of cardinality m for graph G . Each edge $\{v_i, v_j\}$ constitutes a clique of size 2. In the intersection matrix C , the clique (edge) is represented by row k with $C(k, i) = C(k, j) = 1$ and other entries in the row being zero. In general, column indices l in row k where $C(k, l) = 1$ constitutes a clique on vertices v_l of graph G . Thus the ECC problem can be cast as a matrix compression problem.

ECC Matrix Problem *Given $A \in \{0, 1\}^{m \times n}$ determine $A' \in \{0, 1\}^{k \times n}$ with k minimized such that the intersection graphs of A and A' are isomorphic.*

Figure 1a displays a graph on 5 vertices. Figure 1b depicts an intersection matrix representing an edge clique cover of cardinality 7 (number of edges). In the figure a dark dot represents numerical value 1 while a blank entry is a zero. The intersection matrix in Fig. 1c corresponds to an edge clique cover with three cliques. This is also the minimum clique cover for the given graph. To verify that it represents a clique cover, we examine each row of the matrix. Row 1 has dots in columns 1, 3, 4 representing the clique on vertices 1, 3, 4. Row 2 represents the clique on vertices 2, 4, 5 and the remaining edge is covered by row 3.

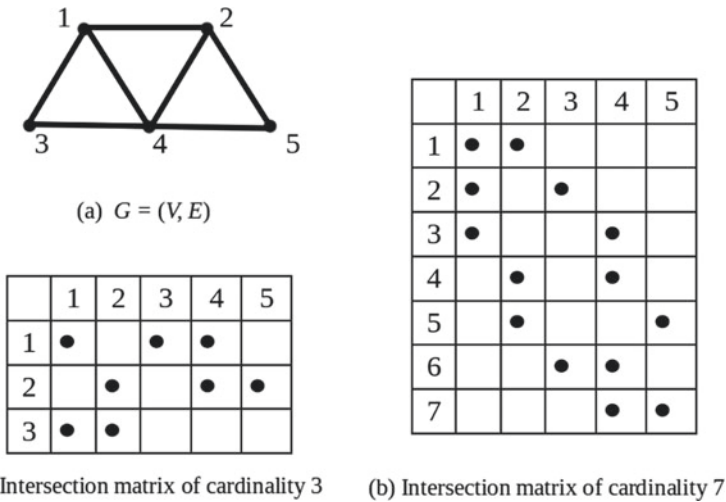


Fig. 1 ECC as a sparse matrix problem

2.2 A Heuristic for Clique Cover

The heuristic algorithm that we have implemented for the ECC problem is based on an algorithm due to Kellerman [11]. For ease of presentation we discuss the algorithm in graph theoretic terms. However, our computer implementation uses sparse matrix framework of DSJM [5] and all computations are expressed in terms of intersection matrices.

There is a close connection between the clique cover of a graph $G = (V, E)$ and the coloring of vertices of the complement graph $\bar{G} = (V, \bar{E})$ where $\bar{E} = \{\{u, v\} \mid \{u, v\} \notin E\}$. In the classical graph coloring problem, vertices of the graph are partitioned into subsets (colors) such that pair of vertices connected by an edge are in different subsets. The optimization version asks for the partition with smallest number of subsets. It is well-known that the greedy coloring heuristic is sensitive to the order in which the vertices are processed (see [3]). Consider an optimal coloring of graph G and order the vertices in nondecreasing color index. It is not difficult to see that the greedy heuristic on graph G with the given order of the vertices produces optimal coloring. We experimentally verify that the ECC heuristic is sensitive to the ordering in which the vertices are processed. We employ three vertex ordering algorithms from the literature: Largest-first order (LFO), Smallest-Last Order (SLO), and Incidence-degree Order (IDO) prior to applying the heuristic [11]. We recall that $d(v) = |N_v|$ denotes the degree of vertex v in graph $G = (V, E)$.

- **(LFO)** Order the vertices such that $\{d(v_i), i = 1, \dots, n\}$ is nonincreasing.
- **(SLO)** Assume that the last $n - k$ vertices $\{v_{k+1} \dots, v_n\}$ in smallest-last order have been determined. The k th vertex in the order is an unordered vertex whose degree in the subgraph induced by

$$V \setminus \{v_{k+1}, \dots, v_n\}$$

is minimum.

- **(IDO)** Assume that the first $k - 1$ vertices $\{v_1, \dots, v_{k-1}\}$ in incidence-degree order have been determined. Choose v_k from among the unordered vertices that has maximum degree in the subgraph induced by

$$\{v_1, \dots, v_k\}$$

Next, we present the algorithm for the ECC problem.

Let the vertices of graph $G = (V, E)$ be ordered in one of SLO, LFO, and IDO: v_1, \dots, v_n . Also, let $V_{\mathcal{P}} = \{v_1, \dots, v_{i-1}\}$ denote the vertices that have been assigned to one or more cliques $\{C_1, \dots, C_{k-1}\}$ and v_i be the vertex currently being processed. Denote by set

$$W = \{v_j \mid j < i \text{ and } \{v_i, v_j\} \in E\}$$

the neighbors of v_i in $V_{\mathcal{P}}$. The task is to assign v_i to one or more of the existing cliques (or create a new clique) such that each edge incident on v_i that connects to a vertex in $V_{\mathcal{P}}$ is covered by a clique. There are three possibilities:

Case I. W is empty: Create a new clique $C_k = \{v_i\}$

Case II. W is not empty:

Case a. There is a clique $C_l, l \in \{1, \dots, k - 1\}$ such that $W = C_l$: add v_i to C_l

Case b. There is no such clique:

i. If $C_l \subset W$ for some l , add v_i to C_l together with uncovered edges from $V_{\mathcal{P}}$. Update W by removing edges that got covered.

ii. If there are uncovered edges after step II(b(i)) create a new clique from an existing clique and add v_i and the incident edges until all the edges of W are covered.

The complete algorithm is presented below.

VertexOrderedECC ($W, list$)

```

1:  $k \leftarrow 0$  ▷ Number of cliques
2: for  $index = 1$  to  $N$  do ▷  $N$  denotes the number of vertices
3:    $i \leftarrow list[index]$  ▷  $list$  contains the vertices in a predefined order
4:   if  $W = \emptyset$  then ▷  $W \leftarrow \{j \mid j < i \text{ and } \{i, j\} \in E\}$ 
5:      $k \leftarrow k + 1$ 
6:      $C_k \leftarrow \{i\}$  ▷  $C_k$  denotes  $k^{th}$  clique
7:   else
8:      $U \leftarrow \emptyset$  ▷ Contains neighbours of  $i$ , which are in the cliques
9:     for  $l = 1$  to  $k$  do
10:      if  $C_l \subseteq W$  then
11:         $C_l \leftarrow C_l \cup \{i\}$ 
12:         $U \leftarrow U \cup C_l$ 
13:      if  $U = W$  then

```



```

14:         break
15:      $W \leftarrow W \setminus U$ 
16:     while  $W \neq \emptyset$  do
17:          $Max \leftarrow \emptyset$ 
18:          $MIN_l \leftarrow 0$ 
19:         for  $l = 1$  to  $k$  do
20:             if  $|Max| < |(C_l \cap W)|$  then
21:                  $Max \leftarrow (C_l \cap W)$ 
22:                  $MIN_l \leftarrow l$ 
23:          $l \leftarrow MIN_l$ 
24:          $k \leftarrow k + 1$ 
25:          $C_k \leftarrow (C_l \cap W) \cup \{i\}$ 
26:          $W \leftarrow W \setminus C_l$ 
27: return  $C_1, C_2, \dots, C_k$ 

```

We argue that the cliques C_1, C_2, \dots, C_k returned by the algorithm constitutes an edge clique cover for the input graph G .

The main **for**-loop (line 2) reads the next vertex (i) from the ordered list of vertices and tries to include it in one of the existing cliques, or creates new clique(s) with vertex i included. If vertex i has no neighbor ($W = \emptyset$) in $V_{\mathcal{P}}$, a new clique gets created (line 6). If the neighbor set W is not empty, the algorithm tries to identify existing cliques C_l that are subsets of W and assigns vertex i to each of them (lines 9 – 15, **Case 2. a.** and **Case 2. b. i.**). This step covers edges of the form $\{i, i'\}$ where $i' \in C_l, C_l \subset W$. Finally, the **while**-loop (line 16) covers the remaining edges (**Case II. b. ii.**) of the form $\{i, i'\}$ where $i' \in S, S = W \cap C'_l, l' \in \{1, 2, \dots, l\}$ with $|S|$ maximum. The maximality on $|S|$ ensures that each newly created clique covers largest number of uncovered edges. For a graph $G = (V, E)$ each edge is a clique of size 2 so that set E constitute an (trivial) ECC. Therefore, each edge of input graph G eventually gets assigned to one of the cliques output by algorithm **VertexOrderedECC**.

The above discussion can be summarized in the following result.

Lemma 1 *The collection $\{C_1, C_2, \dots, C_k\}$ computed by Algorithm **VertexOrderedECC** constitutes an ECC of graph G .*

3 Numerical Testing

In this section, we provide results from numerical experiments on selected test instances. The graph instances are chosen from standard benchmark collections that are used in the literature for ECC and closely related graph problems such as, graph coloring, graph partitioning, etc. The data set for the experiments is obtained from the University of Florida Sparse Matrix Collection [12]. Instances **chesapeake, delaunay_n10 to 13, as-22july06** are from “10th DIMACS Implementation Challenge” benchmark collection for graph clustering and graph partitioning. Instances **ca-GrQc, as-735, Wiki-Vote, p2p-Gnutella04, Oregon-1** are from “Stanford Net-

work Analysis Platform (SNAP)” collection. These instances represent social networks from variety of applications. We also consider the data set for Compact Letter Displays used in [13]. The experiments were performed using a PC with 3.4G Hz Intel Xeon CPU, 8 GB RAM running Linux. The implementation language was C++ and the code was compiled using $-O2$ optimization flag with a g++ version 4.4.7 compiler.

A short description of the data set for our experiments is as follows:

- **chesapeake:** Symmetric, undirected graph and contains 39 vertices and 170 edges.
- **delaunay_n10 to 13:** The graphs are symmetric and undirected. The minimum degree is 3 for all of them and the maximum degrees are 12, 13, 14 and 12 respectively.
- **as-22july06:** The graph is symmetric and undirected having maximum degree 2.4K and minimum degree 1.
- **ca-GrQc:** General Relativity and Quantum Cosmology network covers scientific collaboration between authors in this field. This graph contains an undirected edge from i to j , if author i co-authored a paper with author j .
- **as-735:** An autonomous system which represents a communication network of who-talks-to whom.
- **Wiki-Vote:** This data set contains voting data of Wikipedia till January 2008 where the contest was between volunteers to become one of the administrator. There is a directed edge from node i to node j if user i voted for user j .
- **p2p-Gnutella04:** A snapshot of Gnutella peer-to-peer file sharing network on August 04, 2002. A directed graph where nodes represent hosts and edges represent connection between hosts.
- **Oregon-1:** Undirected graph where autonomous system peering information is inferred from Oregon route-views on May 26, 2001.
- **Triticale, winter wheat and oilseed rape yield trials:** These instances are from the application “compact letter display” [13] to test ECC algorithms.

Test results for the selected test instances from group DIMACS10 and SNAP are reported in Tables 1 and 2 respectively. Test results for Compact Letter Display are reported in Table 3. Here, N represents the number of vertices and M represents the number of edges of the graph. $|C|$ represents number of cliques required to cover all the edges.

For comparison we also show the ECC results where no specific vertex ordering is employed, in addition to ordering algorithms LFO, SLO, and IDO. Column labelled *Natural* reports the ECC result when the vertices are processed in the order they are specified in the data file. On DIMACS10 instances, smallest last order gives the best result except for instance named `as-22july06`. On SNAP instances largest-first order is the overall winner. Note that on both sets of test instances ordered approach produces strictly better ECC compared with *Natural*. We remark that OCaml implementation from [2] fails (hangs) to run on DIMACS10 and SNAP instances. As such no comparison of the ECC quality (size) can be made. Table 3 displays

Table 1 Test results for DIMACS10 matrices

Matrix			Natural	SLO	LFO	IDO
Name	N	M	$ C $	$ C $	$ C $	$ C $
chesapeake	39	170	90	79	83	80
delaunay_n10	1024	3056	1300	1223	1302	1268
delaunay_n11	2048	6127	2610	2482	2617	2527
delaunay_n12	4096	12264	5228	4973	5264	5061
delaunay_n13	8192	24547	10489	9937	10541	10121
as-22july06	22963	48436	34695	34772	34568	34666

Table 2 Test results for SNAP matrices

Matrix			Natural	SLO	LFO	IDO
Name	N	M	$ C $	$ C $	$ C $	$ C $
ca-GrQc	5242	14496	3791	3879	3777	3900
as-735	7716	13895	9055	9108	8985	9038
Wiki-Vote	7115	103689	43497	45530	42482	45491
p2p-Gnutella04	10876	39994	38475	38474	38475	38474
Oregon-1	11174	23409	15736	15807	15631	15857

Table 3 Test results for compact letter displays [13]

Graph			Degree ordered method	Insert- absorb	Clique- growing	Search tree
Name	N	M	$ C $	$ C $	$ C $	$ C $
Triticale 1	13	55	4	4	4	4
Triticale 2	17	86	5	5	5	5
Wheat 1	124	4847	50	56	50	49
Wheat 2	121	4706	48	50	48	48
Wheat 3	97	3559	32	39	32	31
Rapeseed 1	47	576	20	20	20	20
Rapeseed 2	57	1040	20	20	20	20
Rapeseed 3	64	1260	24	24	24	24
Rapeseed 4	62	1085	19	19	19	19
Rapeseed 5	64	1456	19	19	19	19
Rapeseed 6	70	1416	27	27	27	27
Rapeseed 7	74	1758	26	29	27	25
Rapeseed 8	59	1128	17	17	17	17
Rapeseed 9	76	1835	30	30	30	30

results using our degree ordered method and two other algorithms discussed in [13]. `Insert Absorb` and `Search Tree` require exponential running time while `Clique Growing` method is an improved implementation of the heuristic of [11]. `Search Tree` is an exact method that produces optimal ECC. `Degree Order Method` reports the best ECC of our implementation. It is evident from the table that our method produces optimal or near optimal (off by 1) ECC.

4 Conclusion

In this work, we have shown that the connection between large networks and their sparse matrix representation can be exploited to employ efficient techniques from sparse matrix determination literature in graph algorithms [18, 19]. The edge clique cover problem is recast as a sparse matrix determination problem. The notion of *intersection matrix* provides a unified framework that facilitates compact representation of graph data and efficient implementation of graph algorithms. The adjacency matrix representation of a graph can potentially have many nonzero entries since it is the product of an intersection matrix with its transpose. We have shown that, similar to graph vertex coloring problem, the ECC problem is sensitive to ordering of the vertices.

Acknowledgements We thank referees for their many valuable suggestions that helped improve the paper. This research was partially supported by the Natural Sciences and Engineering Research Council (NSERC) under Discovery Grants Program.

References

1. Kepner, J., Gilbert, J.: Graph Algorithms in the Language of Linear Algebra. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2011)
2. Gramm, J., Guo, J., Hüffner, F., Niedermeier, R.: Data reduction and exact algorithms for clique cover. *J. Exp. Algorithmics (JEA)*, **13**, 2–15 (2009)
3. Hossain, S., Khan, A.I.: Exact coloring of sparse matrices. In: Kilgour, D.M., et al. (eds.) *Recent Advances in Mathematical and Statistical Methods*. Springer Proceedings in Mathematics and Statistics, vol. 259, pp. 23–36. Springer Nature, Switzerland AG (2018)
4. Kepner, J., Jananthan, H.: *Mathematics of Big Data: Spreadsheets, Databases, Matrices, and Graphs*. MIT Press (2018)
5. Hasan, M., Hossain, S., Khan, A.I., Mithila, N.H., Suny, A.H.: DSJM: a software toolkit for direct determination of sparse Jacobian matrices. In: Greuel, G.M., Koch, T., Paule, P., Sommese, A. (eds.) *ICMS2016*, pp. 425–434. Springer International Publishing, Switzerland (2016)
6. Hossain, S., Suny, A.H.: Determination of large sparse derivative matrices: structural: orthogonality and structural degeneracy. In: Randerath, B., Röglin, H., Peis, B., Schaudt, O., Schrader, R., Vallentin, F., Weil, V. (eds.) *15th Cologne-Twente Workshop on Graphs & Combinatorial Optimization*, pp. 83–87. Cologne, Germany (2017)
7. Kou, L.T., Stockmeyer, L.J., Wong, C.K.: Covering edges by cliques with regard to keyword conflicts and intersection graphs. *Commun. ACM*, **21**(2), 135–139 (1978)

8. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press (1994)
9. James, O.: Contentment in graph theory: covering graphs with cliques. *Indagationes Mathematicae (Proceedings)*, vol. 80, no. 5. North-Holland (1977)
10. Gramm, J., Guo, J., Hüffner, F., Niedermeier, R.: Data reduction, exact and heuristic algorithms for clique cover. In: *Proceedings of the Eighth Workshop on Algorithm Engineering and Experiments (ALENEX)*. pp. 86–94. SIAM (2006)
11. Kellerman, E.: Determination of keyword conflict. *IBM Tech. Discl. Bull.* **16**(2), 544–546 (1973)
12. SuiteSparse Matrix Collection. <https://sparse.tamu.edu/>. Accessed 02 Oct 2019
13. Gramm, J., Guo, J., Hüffner, F., Niedermeier, R., Piepho, H., Schmid, R.: Algorithms for compact letter displays: comparison and evaluation. *Comput. Stat. Data Anal.* **52**, 725–736 (2007)
14. Nestrud, M.A., Ennis, J.M., Fayle, C.M., Ennis, D.M., Lawless, H.T.: Validating a graph theoretic screening approach to food item combinations. *J. Sens. Stud.* **26**(5), 331–338 (2011)
15. Blanchette, M., Kim, E., Vetta, A.: Clique cover on sparse networks. In: *2012 Proceedings of the Fourteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pp. 93–102. Society for Industrial and Applied Mathematics, 2012 Jan 16
16. Ennis, J.M., Ennis, D.M.: Efficient representation of pairwise sensory information. *IFPress* **15**(3), 3–4 (2012)
17. Tinhofer, G.: Generating graphs uniformly at random. In: Tinhofer, G., Mayr, E., Noltemeier, H., Syslo, M.M. (eds.) *Computational Graph Theory. Computing Supplementum*, vol. 7, pp. 235–255. Springer, Vienna (1990)
18. Hossain, S., Steihaug, T.: Graph models and their efficient implementation for sparse Jacobian matrix determination. *Discrete Appl. Math.* **161**(12), 1747–1754 (2013)
19. Hossain, S., Steihaug, T.: Optimal direct determination of sparse Jacobian matrices. *Optim. Methods Softw.* (2012). <https://doi.org/10.1080/10556788.2012.693927>

Comparing Regularization Techniques Applied to a Perceptron



Bryson Boreland, Herb Kunze, and Kimberly M. Levere

Abstract Overfitting is a common problem that is faced when dealing with neural networks, especially as computers continue to get more powerful, and we have the capability to train larger networks with many free parameters. As a result there is a pressing need to develop and explore different techniques to reduce overfitting; we explore the impact of different regularization terms, and their combinations, in the training phase of a single-perceptron neural network.

Keywords Perceptron · Machine learning · Regularization · Overfitting · Neural network

1 Introduction

Current research [4] suggests that regularization can help to avoid overfitting a neural network and improve how it handles new and unobserved data. For example, [5] illustrates that an ℓ_0 regularization approach smoothes a network model and accelerates its training. Such an approach can be extended to other types of networks beyond traditional Artificial Neural Networks (ANN) such as Convolutional Neural Networks (CNN) and Interval Neural Networks (IANN), which are both important tools in deep learning. In this paper, we will discuss the single perceptron model and perform a quantitative comparison of two commonly used regularization techniques, ℓ_1 and ℓ_2 .

B. Boreland (✉) · H. Kunze · K. M. Levere
University of Guelph, Guelph, Canada
e-mail: bborelan@uoguelph.ca

H. Kunze
e-mail: hkunze@uoguelph.ca

K. M. Levere
e-mail: klevere@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_12

2 Background

In the 1950s and 1960s, Frank Rosenblatt developed the perceptron model [2], inspired by early work done by McCulloch and Pitts [1]. In this section, we present some background information and notation on perceptrons. We note that there are more modern models that are used in present day, such as the commonly used sigmoid neuron, but in this paper we stay focused on the perceptron.

2.1 The Simple Perceptron Model

The simple perceptron model is made up of a single neuron containing two layers, an input layer and an output layer, that maps a specified number of inputs into a single output. The main objective of the perceptron is to classify data that can be separated into two different classifications. We can formally define the model using vector notation.

Definition 1 Let $x = [x_1, \dots, x_n]^T$ be a vector of inputs, $w = [w_1, \dots, w_n]^T$ be a vector of weights, and b be the bias (or threshold). Then the output, y , of a perceptron model can be given by,

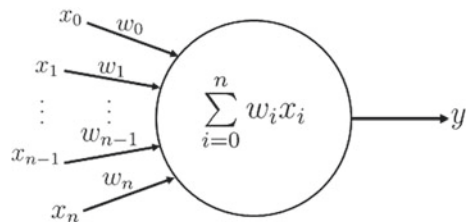
$$y = \begin{cases} 1, & \text{if } w^T x = \sum_{i=1}^n w_i x_i \geq b \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $n \in \mathbb{Z}^+$ is the specified number of inputs. The weights and bias of a single perceptron model can be either boolean or real-valued and can only be used to solve linearly separable problems. Figure 1 illustrates a simple perceptron model.

Remark 1 We typically associate the input $x_0 = 1$ with the bias value b and rewrite b as w_0 . Equation (1) then becomes,

$$y = \begin{cases} 1, & \text{if } w^T x = \sum_{i=0}^n w_i x_i \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Fig. 1 A visualization of the simple perceptron model



In order to calculate the weights and biases efficiently we need an algorithm that updates the new values and a way to quantify if our algorithm is finding values that are optimal. The following section will explore these ideas.

3 Finding Weights and Biases

This section will establish an algorithm to help find the “best” choice (the outputs that are achieved are close to the expected outputs) of weights and define a way to measure the goodness of a particular choice of weights.

3.1 The Loss Function

If we provide our perceptron with a set of inputs and outputs or a “training” set, then we want to find an algorithm that can adjust our weights and bias based on the inputs so that we get the expected output. When the weights of our perceptron have been initialized, we can begin calculating our outputs with the given inputs but the outputs will not be as expected.

A cost function gives a measure of how good our choice of weights and bias are, given our training set and expected outputs. One way to compute an overall cost is to use the sum of squared errors (SSE),

$$C = \sum_i (y_i - \hat{y}_i)^2, \quad (3)$$

where i is a training example, y_i is the expected outcome for the selected training example i , and \hat{y}_i is the predicted outcome based on the current choice of weights and bias. In this paper, we restrict our discussion to convex functions similar to the function introduced in (3).

Our cost is a function of two types of variables, our inputs and our weights on our perceptron. Since we do not have a choice of what our data inputs are, we minimize the cost by changing our weights. The process of seeking weights that minimize the cost function is referred to as “training” the perceptron. Of course, we could use a brute force method in order to find the best solution of weights for our perceptron, but with complexity and increased dimensionality come very expensive computations. Instead we can use gradient descent.

3.2 Gradient Descent

There are a few optimization techniques that can be used to minimize the cost functional and choose optimal weights. One such method is known as gradient descent which is an iterative process that takes steps in a descending direction as defined by the negative of the gradient.

Assume we have a perceptron with weights w_i , $i = 1, \dots, n$, and we want to find which direction to move the weight values in order to minimize a cost function C . We know that the rate of change of the cost function with respect to the weights is as follows,

$$\begin{aligned}\Delta C &= \frac{\partial C}{\partial w_1} \Delta w_1 + \dots + \frac{\partial C}{\partial w_n} \Delta w_n \\ &= \nabla C \cdot \Delta w,\end{aligned}\tag{4}$$

where $\nabla C = \left(\frac{\partial C}{\partial w_1}, \dots, \frac{\partial C}{\partial w_n} \right)$ and $\Delta w = (\Delta w_1, \dots, \Delta w_n)$. Using equation (4) we can now make a choice for Δw that decreases ∇C . We choose,

$$\Delta w = -\eta \nabla C,$$

where the learning rate (or step size) $\eta > 0$ is small. To see why we choose this value, notice

$$\begin{aligned}\Delta C &= \nabla C \cdot \Delta w \\ &= \nabla C \cdot (-\eta \nabla C) \\ &= -\eta \|\nabla C\|^2 \\ &\leq 0\end{aligned}$$

since $\eta > 0$ and $\|\nabla C\|^2 \geq 0$. We use this choice of Δw to create the following update rule for the perceptron's weights

$$w'_i = w_i - \eta \nabla C.\tag{5}$$

If the cost function is non-convex, then the gradient descent algorithm could stop at local minimums rather than the global minimum. In the exploration of this paper, we have chosen a quadratic and therefore convex cost function.

4 Overfitting and Regularization

Modern day neural networks often have a large number of parameters (weights and biases), which can cause a well-known problem of overfitting. In this section we discuss how to combat overfitting by using a technique called regularization. Before

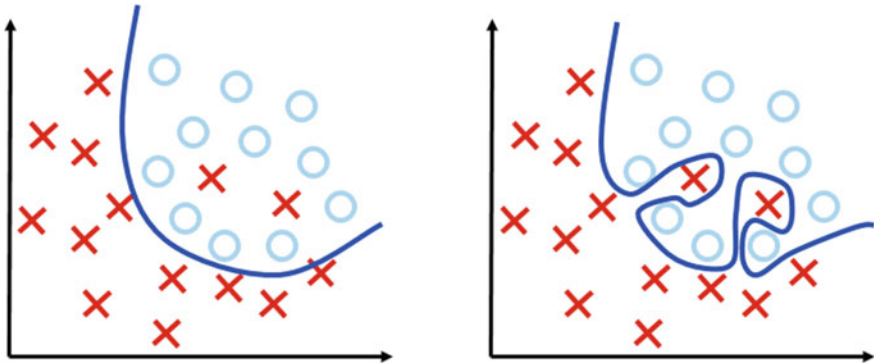


Fig. 2 Example of an appropriately fitted model and an overfitted model

explaining how regularization affects the cost function, a visual example will be introduced to help with understanding what regularization does to our model.

Suppose we have two classes of data that our network is trying to classify, X 's and O 's, by fitting a curve in $2D$ space. Overfitting is when our neural network attempts to fit a line to include every single data point of X on one side of the line and none of the O data points. This can be seen on the righthand side of Fig. 2. The reason why this fit is not appropriate is while our model is trying to find the important properties in the training dataset, it is finding every single possible property. Thus, the application of the resulting neural network to a general dataset can give a poor prediction.

Now, switching to the lefthand side of Fig. 2 we see that an appropriately fitted model generalizes the data well. This sort of outcome is the desired result from using regularization. Mathematically, recalling our SSE cost function (3), we can give a general formula for our new cost function,

$$C_r = C + \lambda \cdot R \tag{6}$$

where R is a vector of regularization terms of choice and $\lambda > 0$ is the regularization coefficient vector that determines the weights of each regularization term. Of course, there are different regularization terms that can be added to the cost function which all have unique properties.

In this work, we consider the use of ℓ_1 and/or ℓ_2 regularization terms, giving

$$\begin{aligned} C_r &= C + \lambda \cdot \mathbf{R} \\ &= C + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \lambda_1 \sum_i |w_i| + \lambda_2 \sum_i w_i^2, \end{aligned} \tag{7}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the regularization coefficients for the ℓ_1 and ℓ_2 terms, respectively. Additionally, we choose to normalize the coefficients by multiplying (7) by $\frac{1}{1+\lambda_1+\lambda_2}$ to get

$$\bar{C}_r = \gamma \cdot C + \gamma_1 \cdot \|w\|_1 + \gamma_2 \cdot \|w\|_2^2,$$

where $\gamma > 0$, $\gamma_1, \gamma_2 \geq 0$, and $\gamma + \gamma_1 + \gamma_2 = 1$. When $\gamma_i = 0$, we do not employ ℓ_i regularization, $i = 1, 2$.

We comment that, as the word implies, regularization is typically employed to make an ill-behaved objective functions more regular. In this work, our initial cost function C is extremely well-behaved, so we should view the regularization terms instead as perturbations that we hope can improve the training of the network.

5 Results and Discussion

We consider the MNIST dataset, introduced by [3], consisting of images that are scanned handwriting samples from 250 people, where half were US Census Bureau employees, and the other half were high school students. The images are greyscale and 28 by 28 pixels in size. The perceptron takes 784 pixels as 784 inputs and then has a single output classifying the image as either a 1 or a 0. An example of the data set can be seen in Fig. 3.

We are interested in how the value of the cost function C changes as we adjust the values of γ_1 and γ_2 in \bar{C}_r for the weight update rule.

When $\gamma_1 > 0$ and $\gamma_2 = 0$ we see from Fig. 4a that with a small amount of γ_1 added to the cost function our error, or value of our cost function, is 0. As we increase γ_1 to add more weight to our ℓ_1 term, the error value grows which suggests that too much γ_1 over penalizes the weights when updated.

On the other hand, when $\gamma_2 > 0$ and $\gamma_1 = 0$ we see from Fig. 4b that we have more choices of our γ_2 value while still keeping the error value at 0. A similar result occurs if we add too much γ_2 and the error value begins to grow. This result can be expected by considering the behaviours of ℓ_1 and ℓ_2 . ℓ_1 decreases linearly as you move towards the origin which will cause our update rule to send the value of the weights to 0. However, ℓ_2 decreases quadratically as you move towards the origin and therefore sends the value of the weights close to zero but not equal to zero.

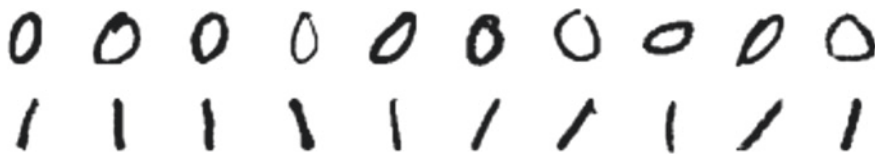


Fig. 3 Examples of the handwritten 0's and 1's included in the MNIST data set

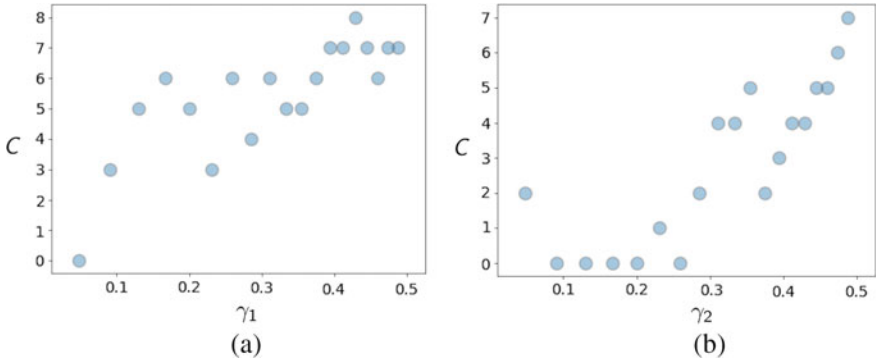


Fig. 4 Value of C for **a** choices of γ_1 (ℓ_1 term) when $\gamma_2 = 0$ (ℓ_2 term) and **b** choices of γ_2 (ℓ_2 term) when $\gamma_1 = 0$ (ℓ_1 term)

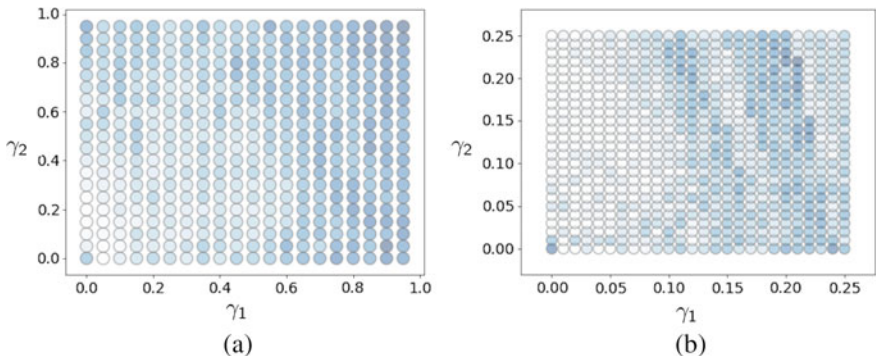


Fig. 5 Visualizations of values of the cost function C for **a** choices of γ_1 and γ_2 and **b** the zoomed section where $0 \leq \gamma_1, \gamma_2 \leq 0.25$. The darker the shade of blue the larger the value of C

When considering both ℓ_1 and ℓ_2 terms present in our weight update rule we get the results we might expect to see based on the individual results. In Fig. 5a we see that when both terms are present adding a small amount of each is when the error value remains low (circles with a lighter shade of blue) or at 0 (white circles). There are once again more choices of γ_2 that still have a positive effect on our model and this is clear when zooming in on our choices of γ_1 and γ_2 as shown in Fig. 5b.

In conclusion, adding a small amount of each term provides a positive effect on our model during training and optimization of the weights. Further investigation needs to be done on the choice of cost function such as an exponential cost or the cross-entropy cost function. Of course, different choices of the cost function can add more layers of complexity to the model that will interact differently with the ℓ_1 and ℓ_2 terms.

The next step in this work is to extend the exploration to a network of many perceptrons.

References

1. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943)
2. Rosenblatt, F.: Principles of neurodynamics: perceptrons and the theory of brain mechanisms. In: Palm, G., Aertsen, A. (eds.) *Brain Theory*. Springer, Berlin, Heidelberg (1962)
3. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Back-propagation applied to handwritten zip code recognition. *Neural Comput.* **1**(4), 541–551 (1989)
4. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: *International Conference on Learning Representations, ICLR 2017* (2017)
5. Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through L_0 regularization. In: *International Conference on Learning Representations, ICLR 2018* (2018)

Key Performance Indicators and Individual Factors on Penalty Kicks



Joao Fialho

Abstract In the recent years quite a few papers have been dedicated to the study of penalty kicks in soccer. With either the intent of predicting the direction or the final outcome of the kick, several different factors have been analyzed, from kinematics, biomechanics, stress levels, individual skills and fatigue as just some examples. In this paper, the author studies a group of four different international soccer players with the objective of identifying key performance indicators on kicks from the penalty spot. Using data analysis techniques, with emphasis on Cramer's V correlation and hypothesis testing, several variables are analyzed, with the intent of identifying global and individual factors, that might provide a signal for which side of the goal post the penalty kick will be aimed at. This study's primary objective is then to provide the goalkeeper with some attributable information that can be used in his advantage, to predict the side for where the penalty is more likely to be aimed at.

Keywords Penalty kicks · Sports analytics · Individual performance indicators in football

1 Introduction

When analyzing a soccer match, one can not exclude the penalty kick. Even more when one specifically considers World Class tournaments, such as the World Cup, European Cup, Champions League, or other international competitions, the penalty kick becomes even more important, as the knockout stages or and even the trophy are sometimes decided on penalty shootouts.

According to informal statistics collected by ESPN, the current rate of conversion of penalty kicks, ranges from 70 to 80%, depending on the League or tournament

J. Fialho (✉)
British University of Vietnam, Hanoi, Vietnam
e-mail: joao.f@buv.edu.vn

Centro de Investigacao em Matematica e Aplicacoes (CIMA-UE),
University of Evora, Evora, Portugal

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_13

being played. In addition to this high conversion rate, it takes on average 400 ms for a ball to reach the top corner, with an average shot speed of 113 km/h, as shown in [3]. The goalkeeper needs 100 ms to process plus 100 ms to decide and initiate movement, and then needs 700 ms to jump and try to reach the ball.

With this information, it becomes clear that, currently, the advantage is on the kicker side, not on the goalkeeper, whether we are looking at response time or even at the current estimated efficiency percentage. It becomes important to counteract this advantage. In order to do so, several papers and authors approach the penalty kick in soccer using several different paradigms, such as biomechanics and kinematics [6–8, 10] or psychology [2, 5, 11], as examples. However as they aim for a generalization, they do not provide the goalkeeper with any clear guidance on how to reduce the kicker advantage.

In this paper the author proposes an individual approach to each penalty taker, in an attempt to identify key performance factors. The *a priori* identification of such factors, would clearly provide the goalkeeper with some ability to predict the side for where the shot is more likely to be taken and therefore “even the odds” in a penalty kick.

Two different case studies will be presented, one at a global level covering 2 former players and 2 current players from the Portuguese National squad, and then an individual analysis of two of those players, to better understand the level of specific factors. In both cases, key factors will be exhibited and proposed. The author would like to emphasize on the case study characteristic of this research. Given the fact the the sample is somewhat limited, the factors and testing should not to be extrapolated to be a bigger population than the one in consideration.

2 Data Set and Variables Definition

For this particular case study a total of 176 penalty kicks were analyzed from 4 international professional soccer players from the Portuguese National squad. Two are former players and 2 are current players. The rationale to select the four players was based on the following: three were the players in the Portuguese National team with more penalty kicks taken, in official competitions at the time of the World Cup 2018 and the other player (Player 3 in Table 1) was a youth player. As some of the players are still currently playing for their teams, their identity has been kept confidential. Data recorded covers official league and cup games as well as international club and national team competitions, from 16 different competitions. For these four players, this represent their totality of penalty kicks in official competitions, from the season 2005/2006 to May 2018.

Data was analyzed and compiled from video recordings of each penalty shot, a second observer confirmed the observations by taking a random sample of the initial observations. For each penalty kick, 17 different variables were analyzed. The distribution in terms of penalty kicks taken, over the four players, is not even. Individual distribution is given in the Table 1.

Table 1 Players and number of penalty kicks considered

Player	Number of penalty kicks
Player 1	122
Player 2	26
Player 3	9
Player 4	19

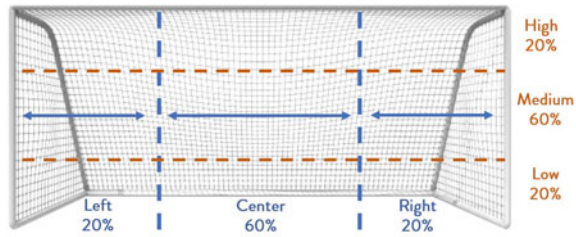
Table 2 Variables considered

Variable	Details	Data type
Side for which GK dive	Right, Center, Left	Nominal
Penalty scored	Yes, No	Nominal
If not, saved or missed	Saved, Missed	Nominal
Shot side	Right, Center, Left	Nominal
Looked at the side before the shot	Yes, No	Nominal
Player faked	Yes, No	Nominal
Step count	Number of steps	Numerical
Shooting technique	Inner part, Front part	Nominal
Shooting type	Skill, Power	Nominal
Shot height	Low, Medium, High	Nominal
Shot speed	Low, Medium, High	Nominal
GK faked	Yes, No	Nominal
GK stayed with open arms	Yes, No	Nominal
GK looked for visual contact	Yes, No	Nominal
Supporting fans location	Behind the goal, Opposite side	Nominal
Moment of the game	0 to 15 min, 16 to 30 min, 31 to 45 min, 46 to 60 min, 61 to 75 min, 76 to 90 min, Over 90 min	Nominal
Result at the time	Winning, Drawing, Losing	Nominal

In terms of variables collected, they are listed in Table 2, along with the details

The variables collected are in line with variables suggested in [2, 5, 7–9] and cover kinematic and psychological observable factors. As mentioned previously, one of the main objectives in this study was to concentrate on variables that could be assessed and inferred by the goalkeeper during a match. Variables such as Speed and Shot height cannot be observed before the penalty is taken, but they provide a more in depth analysis of each of the penalty takers.

Fig. 1 Considerations in terms of side and height



2.1 Variable Clarification

Some of the variables defined require some more clarification and detail.

As specified in [8], due to goalkeeper's position, the author considers center as the central 60% portion of the goal, the left side and right side cover the remainder, as illustrated in Fig. 1.

A similar approach was made for the height. The "medium height" area was considered as the region from the goalkeeper's knee up to 80% of full arm extension region, as suggested in [8].

In terms of speed, the split was made based on the information cited in [3]. Therefore, shots that took between 350 and 450 ms (inclusive) to reach the goal were considered medium speed, less than 350 ms high speed and more than 450 ms, low speed. This was analyzed by an approximate measure from the time the shot was taken until it crossed the goal line. A second observer confirmed a sample of the observations.

As per the variable "Player faked", the binary input Yes/No refers to the deceiving action of the player. If the player slows down or tries to deceive the goalkeeper, during his run to the ball, before taking the shot, that action is recorded as Yes. If the player does not attempt such actions, it is recorded as No. The variables collected that are related to goalkeeper behavior (GK faked, GK stayed with open arms, GK looked for visual contact) were included to understand if the GK behavior could have a significant influence in the choice of side selected by the penalty taker. In terms of the player variables, the binary input Yes/No was used. The "Moment of the game" variable was split in 15 min intervals. This split mimics the influence of both stress and fatigue levels, as mentioned in [4, 5]. The "Result at the time" and "Supporting fans location" are recorded, with the intent to measure the external pressure on the penalty kick taker.

3 Methodology

The approach taken in this paper focuses mainly on identifying key performance factors, first at a more global level for group of four players, and then uses the same process to analyze one individual player to assess potential indicators that can provide some insight to a goalkeeper, during a penalty situation.

The techniques used rely on the measure of association Cramer's V, as suggested in [1], combined with hypothesis testing. The reason for such is related to the fact that the main variables under scrutiny are nominal variables.

For the global approach, a series of factors will be analyzed and some hypothesis testing will be done to study the independence of the factors under consideration. In terms of the the individual player analysis, the author will identify overall accuracy, current tendency and efficiency, key performance indicators and then present a probability cross tabulation table, emphasizing the player's tendency, based on the main factors previously identified. Hypothesis testing on the factors will be conducted for each individual player as well. A level of significance of 5% was considered.

These dependent factors, can then be seen as a predictive model for each players choice of side, under those specific conditions.

4 Key Performance Factors in Penalty Kick

As mentioned in the introduction to this study, four different international soccer players are analyzed. Combined, a total of 176 penalty kicks were analyzed.

4.1 Global Analysis

As mentioned previously, the penalty kick analyzed represent the universe of all penalty kicks taken, in official games (domestic or international competitions) for these four players. The overall level of efficiency in this sample is 65%, meaning that globally 65% of the penalties resulted in goal.

To identify the key performance indicators, Cramer's V correlation coefficients were calculated for every combination of variables. The resultant graph is shown in Fig. 2.

From Fig. 2, it is clear that the most relevant factors that influence the choice of side are, "Player faked", with a coefficient of 0.34, "Moment of the game", with a coefficient of 0.19 and the "Shooting Type", with a coefficient of 0.19. In terms of analysis of Cramer's V coefficient, these values show that "Player faked" has a strong connection with the choice of side, where "Shooting type" and "Moment of the game", seem to have a weak to moderate connection. A Chi-squared test of independence was ran on the above mentioned factors, at a level of significance of 5%. Results are detailed in Table 3.

Table 3 clearly shows that none of the potential connections is statistically significant. However this analysis is a global one. In the next section a more individual approach is taken.

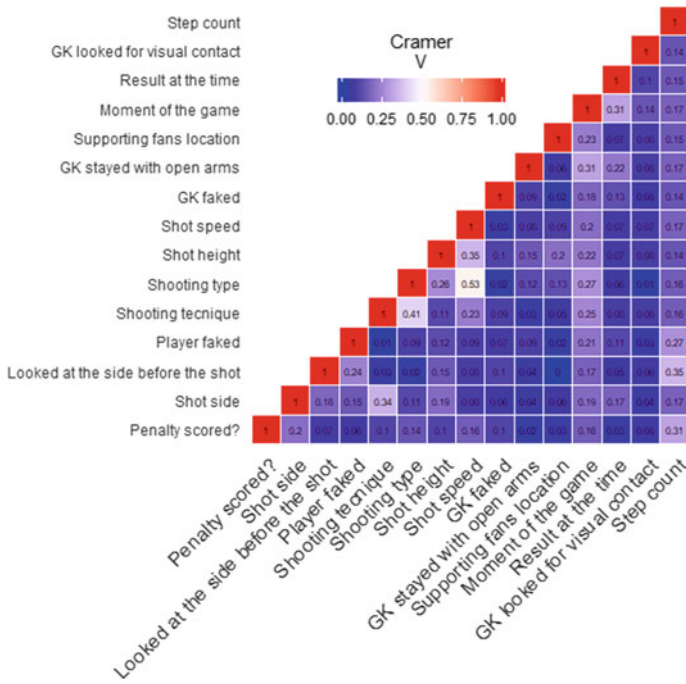


Fig. 2 Heatmap with Cramer’s V coefficients—global analysis

Table 3 Chi-squared test of independence

Hypothesis	Chi-squared	p-value	Outcome
H0: Choice of side is independent of Player faked	3.5601	0.1686	Do not reject H0
H0: Choice of side is independent of Moment of the game	12.326	0.7213	Do not reject H0
H0: Choice of side is independent of Shooting type	1.983	0.371	Do not reject H0

4.2 Individual Players

The individual player selected has a total of 122 official penalties taken in 13 different international or domestic competitions. This player shoots preferably with his right foot and all the penalties considered were shot with the right foot. His current level of efficiency is 82%, meaning he successfully scored 100 of the 126 penalties taken.

The player’s preference and efficiency are shown on Fig. 3.

From Fig. 3 it is also clear that there is dominant choice in terms of the left side, as it is chosen 55% of the times. However shots taken to the right side, even though they are less frequent they occur 35% of the time, seem to be more successful (91%

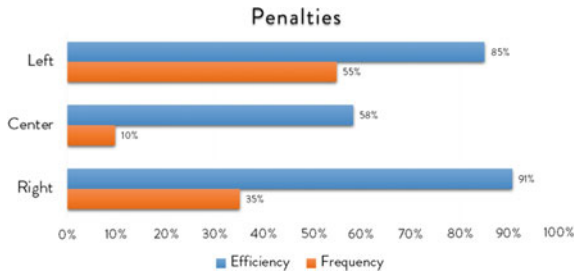


Fig. 3 Graph illustrating tendency and efficiency by player’s choice of side for Player 2

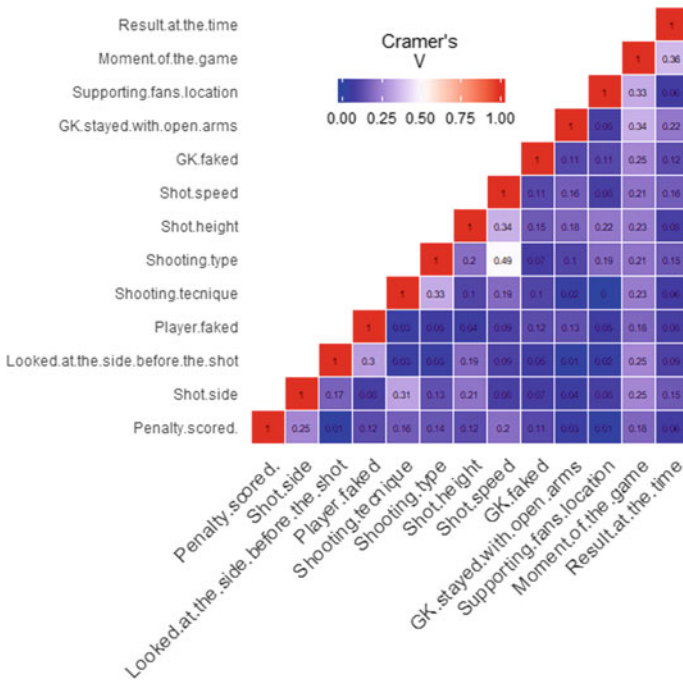


Fig. 4 Heatmap with Cramer’s V coefficients for Player 1

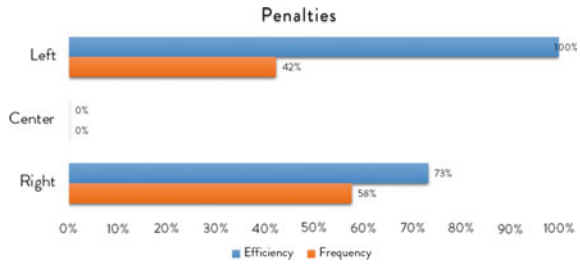
of the times). To identify the key performance indicators, Cramer’s V coefficients were calculated for every combination of variables. The resultant graph is shown in Fig. 4.

From the analysis of Fig. 4, one can identify as most relevant factors for the choice of side, the “Shooting technique”, with a coefficient of 0.31 and the “Moment of the game” with a coefficient of 0.25. These are considered to be strong to moderately strong factors. Results are detailed in Table 4.

Table 4 Chi-Squared test of independence

Hypothesis	Chi-squared	p-value	Outcome
H0: Choice of side is independent of Shooting Technique	11.971	0.002515	Reject H0
H0: Choice of side is independent of Moment of the game	14.868	0.3872	Do not reject H0

Fig. 5 Graph illustrating tendency and efficiency by player’s choice of side for Player 1



From the factors considered only “Shooting technique” is statistically significant. This factor is however harder to identify to the goalkeeper. Nevertheless, looking at the player bio-mechanics, the test shows that there is a statistically significant “give away” of side. Looking at the original data, when the player shoots with the front part of the foot, 70% of the shots will go left.

To highlight the gain in terms of detail when analyzing individual players, another individual player is analyzed using the same process. This player has a total of 26 official penalties taken in 3 different international or domestic competitions. His preferred foot is the right foot and all the shots analyzed in this paper were taken with the right foot. His current efficiency is 85%, as he successfully scored 22 of the 26 penalties taken.

The player’s preference and efficiency are shown on Fig. 5.

Cramer’s V coefficients were calculated for every combination of variables. The resultant graph is shown in Fig. 6.

the most relevant factors for the choice of side are, the “Moment of the game”, with a coefficient of 0.54, the “Shooting speed”, with a coefficient of 0.42 and the “Result at the moment of the shot”, with a coefficient of 0.26. The table with the hypothesis testing is shown below (Table 5).

Showing in this case that the only statistically relevant factor for this player is the Result at the moment of the shot.

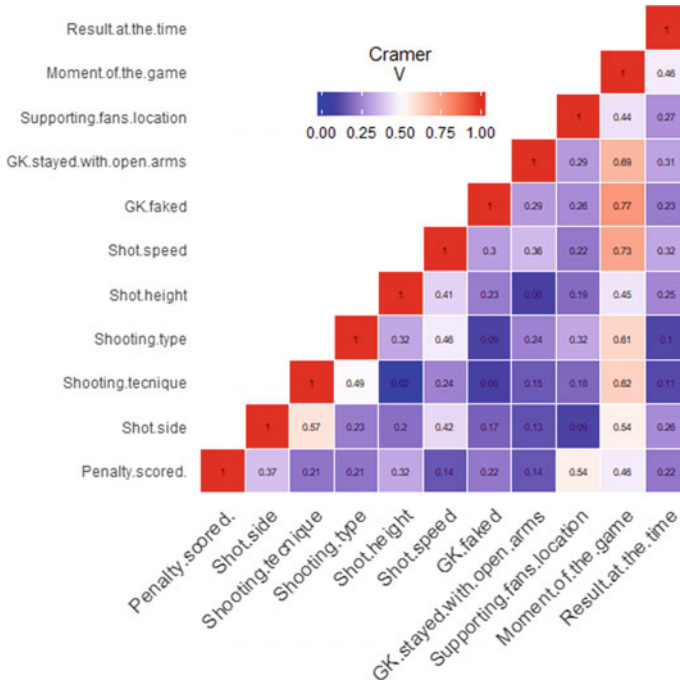


Fig. 6 Heatmap with Cramer’s V coefficients for Player 1

Table 5 Chi-squared test of independence

Hypothesis	Chi-squared	p-value	Outcome
H0: Choice of side is independent of Moment of the game	6.9333	0.4359	Do not reject H0
H0: Choice of side is independent of Shooting speed	4.2386	0.1201	Do not reject H0
H0: Choice of side is independent of Result at the moment of the shot	6.7394	0.0344	Reject H0

5 Conclusion

In an initial approach in this same study, more than 176 penalty kicks were analyzed, from different leagues and competitions and the connections found between variables at a global level were not significant.

In this study the focus is on taking an individual approach to determine key indicators for each player that might “give away” their choice of side, when taking the penalty kick. As the literature shows, these factors can range from biomechanical, to kinematics, to psychological.

In the first part of this study, the generalization, shows that it is very difficult to find common factors on a penalty kick situation, even on a specific team or smaller subset of players. However when the analysis focuses on a specific player, individual characteristics seem to emerge. In both of the cases analyzed it was possible to identify observable factors, that can provide the goalkeeper with some “a priori” information, to offset the player advantage in a penalty kick, in soccer. Predicting the side or outcome of a penalty, in a generalized manner, was not the goal of this study, but more to provide an overview on a technique that can be useful in turning the penalty kick lottery, into a more balanced event. On an individual basis and based on historic player data it was possible not only to identify those factors but also to use them to build an individual predictive model, based on those same factors.

Acknowledgements The author would like to thank Alexandre Real, Rodrigo Silva and Pedro Zorro. Their contribution in terms of data collection and feedback about the variables to collect was essential. For that and all the support provided, my deepest thanks.

References

1. Cramer, H.: *Mathematical Methods of Statistics*. Princeton University Press, Ch. 21 (1946)
2. Diaz, G., Fajen, B., Phillips, F.: Anticipation from biological motion: the goalkeeper problem. *J. Exp. Psychol.: Human Percept. Perform.* **38**(4), 848–864 (2012)
3. Franks, I., Harvey, T.: Cues for goalkeepers. High-tech methods used to measure penalty shot response. *Soccer J.* **42**, 30–33 (1997)
4. Hughes, M., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C.: Moneyball and soccer—an analysis of the key performance indicators of elite male soccer players by position. *J. Hum. Spt. Exe.* **7**, 402–412 (2012)
5. Jordet, G., Hartman, E., Visscher, C., Lemmink, K.A.P.M.: Kicks from the penalty mark in soccer: the roles of stress, skill, and fatigue for kick outcomes. *J. Sports Sci.* **25**(2), 121–129 (2007)
6. Katis, A., et al.: Mechanisms that influence accuracy of the soccer kick. *J. Electromyogr. Kinesiol.* **23**(1), 125–131 (2013)
7. Li, Y., Alexander, M.J.L., Glazebrook, C.M., Leiter, J.: Prediction of Kick Direction from Kinematics during the Soccer Penalty Kick. *IJKSS.* **3**(4), 1–7 (2015)
8. Lopes, J., Jacobs, D., Travieso, D., Araújo, D.: Predicting the lateral direction of deceptive and non-deceptive penalty kicks in soccer from the kinematics of the kicker. *Human Movement Science* **36**, 199–216 (2014)
9. Mackenzie, R., Cushion, C.: Performance analysis in football: a critical review and implications for future research. *J. Sport. Sci.* **31**(6), 39–676 (2013)
10. Scurr, J., Hall, B.: The effects of approach angle on penalty kicking accuracy and kick kinematics with recreational soccer players. *J. Sports Sci. Med.* **8**(2), 230–234 (2009)
11. van der Kamp, J., Dicks, M., Navia, J.A., et al.: Goalkeeping in the soccer penalty kick. *Ger. J. Exerc. Sport Res.* **48**(2), 169–175 (2018)

Sparse Covariance and Precision Random Design Regression



Xi Fang, Steven Winter, and Adam B. Kashlak

Abstract Linear regression for high dimensional data is an inherently challenging problem with many solutions generally involving some structural assumption on the model such as lasso's sparsity in the parameter vector. Considering the random design setting, we apply a different sparsity assumption: sparsity in the covariance or precision matrix of the predictors. Thus, we propose a new regression estimator by first applying methods for estimating a sparse covariance or precision matrix. This matrix is then incorporated into the estimator for the regression parameters. We mainly compare this methodology against the classic ridge or Tikhonov regularization method.

Keywords Graphical lasso · Penalized estimator · Ridge regression · Thresholding

1 Introduction

Linear regression is a backbone of statistical methodology. The classical least squares approach has a simple and elegant theory, but fails in the high dimensional setting where the number of parameters p is greater than the sample size n . High dimensional datasets have led to over 40 years of research resulting in methods such as ridge regression, lasso, elastic net, SCAD, and many others [4]. In this work, we contribute to the compendium of such methods by constructing an estimator for high dimensional regression models making use of sparse covariance and sparse precision matrix estimators in the random design setting.

X. Fang · S. Winter · A. B. Kashlak (✉)
University of Alberta, Edmonton, AB T6G 2G1, Canada
e-mail: kashlak@ualberta.ca

X. Fang
e-mail: xfang@ualberta.ca

S. Winter
e-mail: szwinter@ualberta.ca

© Springer Nature Switzerland AG 2021
D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_14

Sparsity is not new to linear regression. The renowned lasso estimator [18] is one of the most important recent contributions to statistics and mathematics. However, the assumption of lasso is sparsity in the parameters and is used for model selection. In contrast, we consider sparsity in the covariance or precision matrix of the random design matrix X . That is, we assume most off-diagonal entries to be zero and construct a regression estimator under this assumption.

Remark 1 (Sparse Covariance and Precision Matrices) Though being inverses of one another, a sparse covariance and a sparse precision matrix result in two different implications for the underlying data. The covariance matrix considers the marginal correlation between each pair of random variables. A zero entry implies that there is no linear relation between these two variables and in the Gaussian case implies pairwise independence. The precision matrix considers the conditional correlation structure of the data. A zero entry implies that the two variables are uncorrelated—independent when Gaussian—conditioned on the remaining random variables. Thus, the precision matrix defines a network structure and is useful in the study of Gaussian graphical models. Sparse covariance and precision matrices arise in many high dimensional datasets such as genomics, climate, and socioeconomics.

2 Estimator Construction

We begin with a set of n predictor-response pairs (y_i, x_i) , $i = 1, \dots, n$ with $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ and assume that x_{1j}, \dots, x_{nj} for $j = 1, \dots, p$ and the y_1, \dots, y_n are centred as is common in the penalized regression literature. The standard theory of the least squares estimator when $p < n$ for the linear model,

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

with unknown parameters β and mean-zero errors ε_i , yields the *ordinary least squares* (OLS) estimator $\hat{\beta}^{\text{ols}} = (X^T X)^{-1} X^T Y$ for design matrix X with the ij th entry x_{ij} and $Y = (y_1, \dots, y_n)$. Under random design [3, 10], the rows of X are treated as iid random vectors, and the least squares loss $L_{\text{ols}}(\tilde{\beta}) = \mathbb{E} \|Y - X\tilde{\beta}\|_{\ell^2}^2$ is minimized by $\beta = (\mathbb{E}[X^T X])^{-1} \mathbb{E}(X^T Y)$ where we write $\Sigma = n^{-1} \mathbb{E}[X^T X]$, the $p \times p$ covariance matrix for the rows of X . We denote the i th row of X to be x_i .

When $p > n$, the standard covariance estimator $\hat{\Sigma} = n^{-1} \sum_{i=1}^n x_i^T x_i$ is known to be far from Σ and furthermore not full rank and thus not invertible. The classic ridge regression solution—also called Tikhonov regularization—considers the ℓ^2 penalized least squares loss

$$L_R(\tilde{\beta}) = \mathbb{E} \|Y - X\tilde{\beta}\|_{\ell^2}^2 + \lambda \|\tilde{\beta}\|_{\ell^2}^2$$

with solution $\hat{\beta}^R = (X^T X + \lambda I_p)^{-1} X^T Y$ for $p \times p$ identity matrix I_p . This estimator is known to *shrink* the values of $\hat{\beta}^R$ towards zero adding some bias for a sizeable reduction in the estimator's variance. In Sect. 3, we compute the ridge estimator via the `glmnet` package in R [9].

Let $r = \min\{n, p\}$ be the rank of X . The singular value decomposition for the $n \times p$ design matrix can be written as $X = U D V^T$ where D is the $r \times r$ diagonal matrix of singular values, and U and V are $n \times r$ and $p \times r$ matrices, respectively. Thus, for high dimensional data, $r = n$ and U is orthonormal whereas $V^T V = I$ but $V V^T \neq I$ and is, in fact, a projection onto the n -dimensional row space of X , a linear subspace of \mathbb{R}^p . Under this decomposition and the notion of a pseudo-inverse, the OLS estimator can be extended to high dimensional data as

$$\hat{\beta}^{\text{ols}} = V D^{-2} V^T V D U^T Y = V D^{-1} U^T Y.$$

Similarly, the ridge estimator becomes $\hat{\beta}^R = V(D^2 + \lambda I)^{-1} D U^T Y$. Note that the squared singular values d_1^2, \dots, d_n^2 on the diagonal D^2 are the estimated non-zero eigenvalues for the covariance matrix Σ . Hence, this regularization method is augmenting the eigenvalues by adding λ to each to get $d_i^2 + \lambda$ for $i = 1, \dots, n$ and just λ for $i = n + 1, \dots, p$. Therefore, the ridge estimator is, in fact, shrinking the estimated eigenvalues for the precision matrix Σ^{-1} to zero as $\lambda \rightarrow \infty$. This, in turn, takes $\hat{\beta}^R$ to zero.

The ridge estimator replaces $X^T X$ with $X^T X + \lambda I$. Building off of this inspiration, our proposed methodology is to replace $X^T X$ with a sparse covariance estimator, or replace the undefined $(X^T X)^{-1}$ with a sparse precision matrix estimator.

2.1 Replacing $X^T X$

Under the sparsity assumption stated in the introduction for either the covariance or precision matrix, we could consider an alternative estimator being inspired by the Naive-Bayes method. Specifically, we could compute $\hat{\Sigma}^{\text{diag}}$ being the empirical covariance estimator from before with all off-diagonal entries set to zero. Then as this matrix is invertible, we can consider the estimator $\hat{\beta}^{\text{NB}} = (n \hat{\Sigma}^{\text{diag}})^{-1} X^T Y$ where $\hat{\Sigma}^{\text{diag}}$ is multiplied by n to undo the normalization in the covariance estimator. If we were to normalize the data such that $n \hat{\Sigma}^{\text{diag}} = I_p$, then our estimator would be merely $X^T Y$ or equivalently the ℓ^2 inner products between Y and each of the p columns of X .¹ However, the removal of all off-diagonal entries may be too extreme of a methodology. Instead, we relax away from such a diagonal-only estimator by considering sparse estimators for Σ and Σ^{-1} in the following subsections. However, we first take a look at the implications of replacing $X^T X$ with a different positive definite matrix M .

¹ The estimator $X^T Y$ occurs in practice in orthogonal experimental designs when X is chosen such that $X^T X = I_p$ assuming $p < n$. [19].

Let M be a positive definite symmetric matrix with eigen-decomposition WSW^T for W the orthonormal matrix of eigenvectors and S the diagonal matrix of eigenvalues. We wish to write a new regression estimator $\hat{\beta}^M = M^{-1}X^TY$. However, this will initially fail as the first n eigenvectors in W will (most likely) not coincide with the n left singular vectors V of X resulting in a nonsensical estimator. Thus, we have to rotate the entire problem. Let W_n be the $p \times n$ matrix consisting of the first n columns of W , and let S_n be the $n \times n$ diagonal matrix with the n principal eigenvalues of M on the diagonal. Replacing

$$X \Rightarrow Z := XW_n^T \text{ and } \beta \Rightarrow \beta^* := W_nV^T\beta \quad (1)$$

gives the rotated model $Y = X\beta + \varepsilon = Z\beta^* + \varepsilon$. The new estimator making use of M is

$$\hat{\beta}^M = M^{-1}Z^TY = W_nS_n^{-1}DU^TY, \quad (2)$$

which is an estimator for the rotated parameter vector β^* . Note that in Sect. 3, we compare a variety of such estimators in mean squared error. As such transformations as in Eq. 1 are isometries, we can still compare mean squared errors estimated over many random simulations as well as the mean squared prediction error for the forest fire and Arizona crime data.

Remark 2 In the above Eqs. 1 and 2, we could instead consider $p \times p$ orthonormal matrices \tilde{W} and \tilde{V} being the eigenvectors of M and X^TX , respectively. Computationally, the resulting estimator will be equivalent as the eigenvalues corresponding to those $p - n$ additional columns will be zero. Hence, any rotation in those directions will not affect the estimator. The above formulation is more computationally efficient by ignoring these extraneous directions.

2.1.1 Sparse Covariance Estimation

There is a vast literature on sparse covariance matrix estimators for high dimensional data. Two broad approaches are penalized estimators [2, 16] and threshold estimators [1, 5, 17]. The latter methods apply a threshold function entrywise to the off-diagonal entries of the empirical covariance matrix, which effectively sets entries below a specified threshold to zero. A threshold is typically chosen via cross validation. As sample sizes are typically small and cross validation is furthermore computationally expensive, a threshold can also be selected by choosing a suitable individual false positive rate $\alpha \in [0, 1]$ being the probability that an off-diagonal entry is falsely included in the support of the estimator—i.e. the probability that the ij th entry in the estimator is not zero given that $\Sigma_{ij} = 0$ [14]. This α acts as a regularization parameter. Indeed, this estimator allows us to relax away from the above naive-Bayes estimator, which would correspond to $\alpha = 0$, by increasing α to allow for a few off-diagonal entries to be non-zero. Such estimators are computed via the R package `sparseMatEst` [13].

Remark 3 (Positive Definiteness) Even though the empirical covariance estimator is positive semi-definite, a thresholded covariance estimator may no longer be. To rectify this problem, let $\hat{\Sigma}^{(\alpha)}$ be a sparse covariance estimator with false positive rate α , and denote the eigenvalues of $\hat{\Sigma}^{(\alpha)}$ in decreasing order to be $\lambda_1^{(\alpha)} \geq \dots \geq \lambda_n^{(\alpha)}$. Then assuming $\lambda_n^{(\alpha)} < 0$ we add $\{|\lambda_n^{(\alpha)}| + \lambda_1^{(\alpha)}/100\}I_p$ to $\hat{\Sigma}^{(\alpha)}$ to make the new estimator positive definite with a condition number of 100.

2.1.2 Sparse Precision Estimation

The most famous method of sparse precision matrix estimation is the graphical lasso [8], but other regularized estimators also exist [6]. Unlike for covariance matrices, threshold estimation of the precision matrix is more challenging as there is no unbiased estimator for Σ^{-1} threshold. However, [12] applies the same idea of individual false positive rate control by thresholding the debiased glasso estimator of [11]. This precision matrix estimation method is also implemented in the R package `sparseMatEst` [13].

3 Numerical Results

3.1 Simulated Data

In this section, we test the following estimators of the form $\hat{\beta}^M = M^{-1}Z^T Y$ from Eq. 2. For M , we consider threshold based sparse covariance estimators from [14] and the standard ridge estimator. For M^{-1} , we consider threshold based sparse precision estimators from [12] as well as the graphical lasso [8]. To gauge the success of each estimator, we estimate the normalized mean squared error,

$$\text{MSE}(\tilde{\beta}) = \left\| \tilde{\beta} - \beta \right\|_{\ell^2}^2 / \|\beta\|_{\ell^2}^2,$$

over 100 replications for $\beta = (1, \dots, 1)$, the rows of X being iid $\mathcal{N}(0, \Sigma)$ for some sparse Σ discussed below, and $Y = X\beta + \varepsilon$ with iid $\varepsilon_i \sim \mathcal{N}(0, 4)$.

Figure 1 contains the results—estimated log base-2 mean squared errors—for such simulations for Σ tridiagonal with main diagonal 1 and off-diagonal entries 0.4 and contains results for Σ banded with main diagonal 1 and three off-diagonals with values 3/4, 1/2, and 1/4. Since these methods normalize the variance of the predictors before penalizing, we only consider settings where all diagonal entries are 1. For all methods, many choices of the tuning parameter— $\alpha \in [0, 1]$ for sparse covariance and $\lambda \geq 0$ for ridge and glasso—were considered and the best was taken. Hence, Fig. 1 displays results for optimal choice in tuning parameter. In both cases,

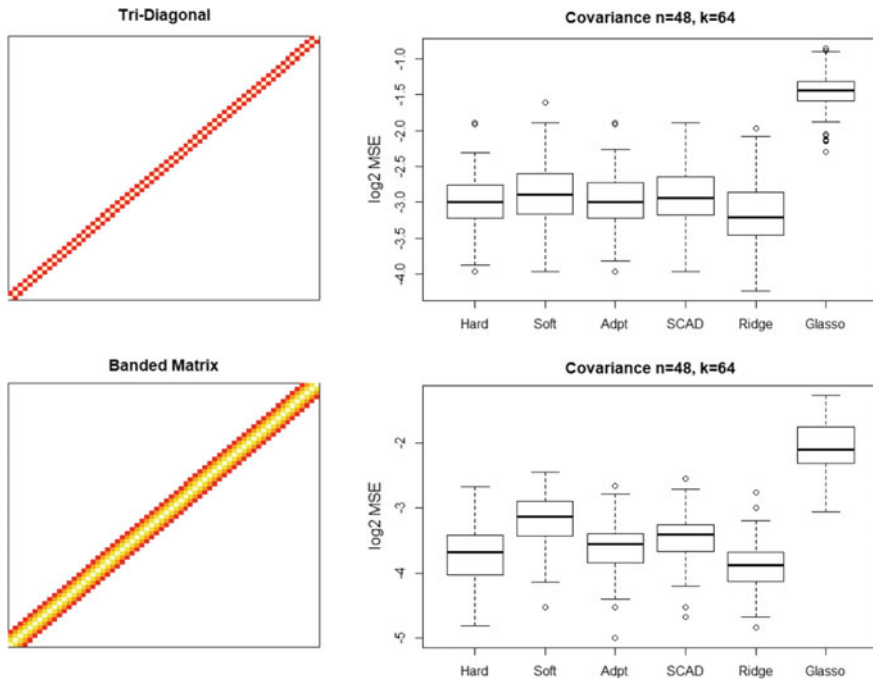


Fig. 1 Top row: Results of methods given a tri-diagonal covariance matrix. Bottom row: Results of methods given a banded covariance matrix. Here, $n = 48$, $p = 64$, and the plots were the result of 100 replications with $\beta = (1, \dots, 1)$

the performance of the sparse covariance methodology was on par with that of ridge regression, and both of these outperformed the graphical lasso.

This methodology was also tested for sparse precision matrices—i.e. rerunning the above simulations but specifying Σ^{-1} to be tri-diagonal or banded as opposed to Σ . In that setting, the sparse precision methodology performed much more poorly than ridge regression. Hence, those results are not included. The answer as to why the sparse matrix-based regression estimator succeeds for covariance matrices but not for precision matrices remains illusive. However, good performance of the precision estimator is observed in Sect. 3.3.

3.2 Forest Fire Data

For a first real data application, we consider the mean squared prediction errors (MSPE) for different regression methods on the Portuguese forest fire data [7] available online on the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>). We aim to predict the log(Area Burned) based on a variety of indices

and weather measurements: the coordinates of the location of the fire, the Fine Fuel Moisture Code (FFMC), the Duff Moisture Code (DMC), the Drought Code (DC), the Initial Spread Index (ISI), the outside temperature, the relative humidity, and the wind speed. For more details, see [7]. Only fires whose total area burned was greater than zero were considered due to the log-transform, which was necessary due to the extreme skewness of the `area` data. Thus, we have $n = 270$ and $p = 9$. Though, this is not high dimensional data, there is strong collinearity among the predictors warranting the use of shrinkage estimators.

To compute the MSPE, we randomly split the data into training and testing sets of sample size $n_{\text{train}} = 225$ and $n_{\text{test}} = 45$, respectively, to fit the model and then compute

$$\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (A_i - \hat{A}_i)^2$$

for A_i the total log-area burned for the i th observation of the test set and \hat{A}_i the i th predicted value. This was averaged over 1000 replications with randomly selected training and testing sets.

Figure 2 displays the results of our methods. The support of the replacement matrix for $X^T X$ is considered on the left for different values of α . Note that for $\alpha = 0.5$, most of the off-diagonal entries have already been removed. The MSPE on the right is considered for sparse covariance matrices with four different types of thresholds: Hard, Soft, Adaptive Lasso, and SCAD thresholding. More details on these can be found in [14, 17]. Ridge regression is also included whereas sparse precision and glasso methods are excluded due to their poor performance on simulated data. Most notably, the methods all return similar MSPE for this dataset, but hard and scad thresholding are the most robust with respect to choice in tuning parameter compared to the other methods.

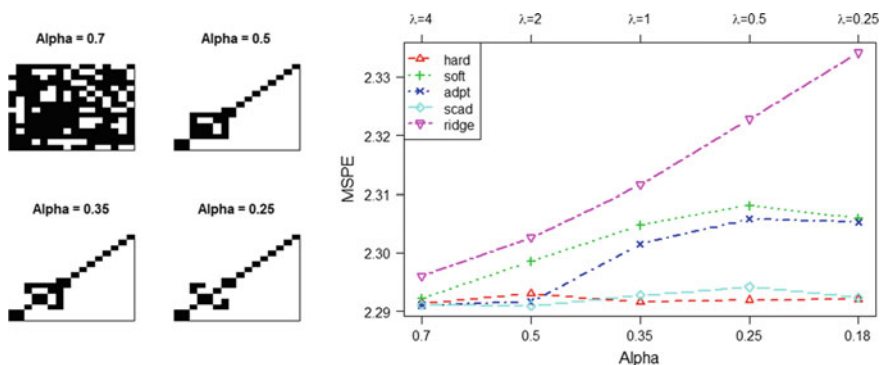


Fig. 2 On the left, the support in black of the thresholded $X^T X$ matrix for different α for the forest fire data [7]. On the right, the mean squared prediction error computed over 1000 replications for 5 different estimators with 5 different values of their respective tuning parameter— α on the bottom for the sparse covariance and λ on the top for ridge regression

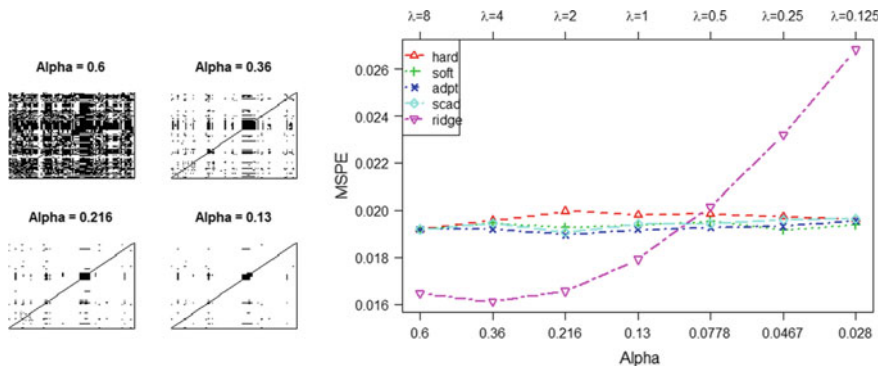


Fig. 3 On the left, the support in black of the thresholded inverse $X^T X$ estimator for different α for the Arizona crime data [15]. On the right, the mean squared prediction error computed over 1000 replications for 7 different estimators with 7 different values of their respective tuning parameter— α on the bottom for the sparse covariance and λ on the top for ridge regression

3.3 Arizona Crime Data

We secondly repeat the previous analysis on the *Communities and Crime Data Set* also from the UCI Machine Learning Repository and discussed in [15]. This dataset contains potential predictors of violent crime collected across the USA. For the sake of our methodology, we only considered the $n = 20$ observations taken from the state of Arizona. There are $p = 99$ predictors in this dataset. The dataset was randomly split into $n_{\text{train}} = 13$ and $n_{\text{test}} = 7$ and the MSPE was computed over 1000 replications.

Figure 3 displays the results for the precision matrix estimator, which performed better than the covariance-based approach. This is reasonable as many predictors may be correlated—e.g., number of homeless shelters and number of vacant houses—but conditionally uncorrelated—e.g., taking median income into account. Here, all precision thresholding methods had very similar MSPE. In contrast, the ridge estimator either performed better or worse depending on choice of λ ; though the scale of the vertical axis indicates that ridge regression only achieves a slightly better MSPE after careful tuning of λ .

4 Discussion

In this article, we proposed an alternative estimator for the parameters of a high dimensional regression model under the random design setting where it is assumed that the rows of the design matrix have a sparse covariance or sparse precision structure. Such structural assumptions do occur in real data problems and are distinct from the usual notation of regression sparsity—that is, sparsity in the parameter vector β .

The result of multiple simulation experiments, beyond what is detailed in Sect. 3, indicate that using a sparse covariance estimator in place of $X^T X$ can achieve similar but no superior results to that of standard ridge regression. Replacing $(X^T X)^{-1}$ with a sparse precision estimator or the classic graphical lasso estimator did not yield good performance in contrast to ridge regression for simulated data. However, we did see strong performance on the Arizona crime dataset.

The success of this methodology does warrant further investigations into such methods considering how they can be improved and if there are scenarios where they can outperform standard ridge regression. Even though the performance of ridge regression was comparable to our methodology, it did not perform significantly better. Also, our method appears more robust to choice of tuning parameter meaning one can achieve similar performance to ridge regression without the need to carefully tune λ .

Acknowledgements The authors would like to thank Dr. Xu (Sunny) Wang from Wilfrid Laurier University and Dr. Yan Yuan from the University of Alberta for organizing the special session on Interdisciplinary Data Analysis of High-Dimensional Multimodal Data at AMMCS 2019 where this work was presented. We would also like to thank the comments of the anonymous reviewers who helped improve this work.

References

1. Bickel, P.J., Levina, E.: Covariance regularization by thresholding. *The Ann. Stat.* 2577–2604 (2008)
2. Bien, J., Tibshirani, R.J.: Sparse estimation of a covariance matrix. *Biometrika* **98**(4), 807–820 (2011)
3. Breiman, L., Freedman, D.: How many variables should be entered in a regression equation? *J. Am. Stat. Assoc.* **78**(381), 131–136 (1983)
4. Bühlmann, P., Van De Geer, S.: *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media (2011)
5. Cai, T., Liu, W.: Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.* **106**(494), 672–684 (2011)
6. Cai, T., Liu, W., Luo, X.: A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* **106**(494), 594–607 (2011)
7. Cortez, P., Morais, A.D.J.R.: A data mining approach to predict forest fires using meteorological data (2007)
8. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2008)
9. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**(1), 1 (2010)
10. Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. In: *Conference on Learning Theory*, pp. 9–1 (2012)
11. Jankova, J., Van De Geer, S.: Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9**(1), 1205–1229 (2015)
12. Kashlak, A.B.: Non-asymptotic error controlled sparse high dimensional precision matrix estimation. *J. Multi. Anal.* **181**, 104690 (2021)
13. Kashlak, A.B.: sparseMatEst: sparse matrix estimation and inference (2019b). <https://CRAN.R-project.org/package=sparseMatEst>. R package version 1.0.0

14. Kashlak, A.B., Kong, L.: Nonasymptotic support recovery for high dimensional sparse covariance matrices. *Stat.* e316 (2020)
15. Redmond, M., Baveja, A.: A data-driven software tool for enabling cooperative information sharing among police departments. *Eur. J. Oper. Res.* **141**(3), 660–678 (2002)
16. Rothman, A.J.: Positive definite estimators of large covariance matrices. *Biometrika* **99**(3), 733–740 (2012)
17. Rothman, A.J., Levina, E., Zhu, J.: Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **104**(485), 177–186 (2009)
18. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc.: Series B (Methodological)* **58**(1), 267–288 (1996)
19. Wu, C.J., Hamada, M.S.: *Experiments: Planning, Analysis, and Optimization*, vol. 552. Wiley (2011)

Applying Neural Networks to a Fractal Inverse Problem



Liam Graham and Matthew Demers

Abstract With the increasing potential of convolutional neural networks in image-related problems, we apply these methods to a fractal inverse problem: Given the attractor of a contractive iterated function system (IFS) what are the parameters that define that IFS? We create and analyze fractal databases, and use them to train various convolutional neural networks to predict these parameters. The neural network outputs produce visually different fractals, however, they could be used to create an initial population for other search algorithms. Additionally, the neural networks become increasingly accurate with increasing numbers of functions defining the IFS.

Keywords Iterated function systems · Applied analysis · Inverse problems · Neural networks · Machine learning

1 Introduction

Fractals have been used to model natural phenomena such as plants and mountains in a wide variety of applications. They are infinite sets that are typically found through fixed point iteration, and contain self-similar features. There are many different types of fractals, one of the most common being those generated by an iterated function system (IFS); a finite set of N contraction mappings on a complete metric space. These mappings possess a unique non-empty compact fixed set of points called the attractor, which is a fractal [2].

Although it is quite easy to generate an image representing the fractal set of an IFS, the inverse problem is quite difficult, and is the focus of this paper. That is, we want to obtain an IFS possessing an attractor closely approximating a given image.

L. Graham (✉) · M. Demers
University of Guelph, Guelph, Canada
e-mail: lgaha07@uoguelph.ca

M. Demers
e-mail: mdemers@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_15

157

For simplicity, we are restricting ourselves to binary (black and white) images, and IFSs consisting solely of affine maps.

Attempts at this problem include using a Gröbner basis [1], wavelet transforms [3], moment matching [11], and genetic algorithms [8]. The most recent approach uses a swarm intelligence method called the cuckoo search, giving satisfactory results for a few specific fractals [9]. However, the problem still remains for general binary images. With the development of computer vision techniques using neural network algorithms, image classification problems have been solved with better than human level accuracy [5, 6, 12]. We will use modified versions of these algorithms and train a convolutional neural network to predict the parameters in an IFS that approximates a given image.

2 Mathematical Background

If $W = \{w_i : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \mid i = 1, 2, \dots, N\}$ is an IFS, then each w_i is a contraction mapping, and W has a unique non-empty compact fixed set given by $S = \bigcup_{i=1}^N w_i(S)$.

For this work, we consider only affine maps in \mathbb{R}^2 . Therefore our functions within the IFS have the form:

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} = \mathbf{A}_i \mathbf{x} + \mathbf{b}. \quad (1)$$

In order to train the neural network, we need data to both give to the neural network (the binary images), and compare with the outputs (the $6N$ parameters shown in Eq. 1, see Sect. 3 for details). We will generate this data by randomly initializing IFSs. However, to ensure that the resulting IFS has an attractor, we must be able to test whether the functions it consists of are contractive. Let k_i be the lipshitz constant for the i th function, w_i , in an IFS. Then,

$$\begin{aligned} \|w_i(\mathbf{x}) - w_i(\mathbf{y})\|_2 &\leq k_i \|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^2 \\ \|\mathbf{A}_i\|_2 \|\mathbf{x} - \mathbf{y}\|_2 &\leq k_i \|\mathbf{x} - \mathbf{y}\|_2 \\ \|\mathbf{A}_i\|_2 &\leq k_i. \end{aligned} \quad (2)$$

Thus w_i is a contraction mapping if $\rho(\mathbf{A}_i \mathbf{A}_i^T) < 1$, where ρ denotes the spectral radius. Since $\mathbf{A}_i \mathbf{A}_i^T$ is positive, semi-definite, and symmetric, all its eigenvalues will be positive. Thus, we can guarantee the eigenvalues will either be larger or smaller than 1 with the restriction $f_i(1) > 0$, where f_i is the characteristic polynomial of $\mathbf{A}_i \mathbf{A}_i^T$. Imposing the condition $\|\mathbf{A}_i \mathbf{x}\|_2 < \|\mathbf{x}\|_2$ further forces both eigenvalues to be smaller than 1. Applying the standard basis for \mathbb{R}^2 to this last restriction, we obtain the following conditions for w_i to be a contraction mapping:

$$\begin{cases} \text{Tr}(\mathbf{A}_i \mathbf{A}_i^T) - \det(\mathbf{A}_i)^2 < 1 \\ a_i^2 + c_i^2 < 1 \\ b_i^2 + d_i^2 < 1 \end{cases} \quad (3)$$

To generate the data, we then initialize all of the IFSs parameters with double precision numbers randomly chosen between -1 and 1 , and check these conditions. If they are satisfied, the image representing the IFSs attractor is generated, if not, the parameters are regenerated. In order to ensure that the value of a pixel is consistent throughout the data, a viewing window spanning from -1 to 1 in both the horizontal and vertical directions was chosen. If the generated fractal does not fit within this window, the largest in magnitude x or y coordinate of a point in the approximated fractal set is used as a divisor of all p_i and q_i parameters in the IFS. Upon a new generation of this set, the attractor will then be scaled to fit within the desired region.

3 Neural Network Background

Neural networks are connections of nodes, which can be thought of as placeholders for values, used to approximate complicated functions [4]. These nodes are organized in layers with the first layer being the input(s), the last layer being the output(s), and all layers in between being hidden layers. As data is fed through the connections, it gets multiplied by values known as weights and is typically transformed by non-linear functions, among other things [4]. Once the output is obtained from the network, it is compared with known values and an error is calculated by a loss function [4]. This error is used to update parameters of the network using modified gradient descent algorithms, resulting in more desirable outputs [4].

Convolutional neural networks (CNNs) are a specific class of neural network commonly used in image related problems. As opposed to fully connected neural networks where all nodes of one layer are connected to all nodes of the next, CNNs use a filter which passes over the image and computes a discrete convolution given by

$$F(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n), \quad (4)$$

where i and j are pixel locations on an image, I , and K is the filter which m and n iterate over [4]. The output of the convolution is then stored in a node of the next layer, effectively creating a feature map. As the network parameters are updated, the filter learns features important to the data-set on which it is trained [4]. Filters in layers near the input may learn to detect lines and edges whereas filters further in the network learn to detect objects constructed by those lines [4].

These convolutions are necessary in image related problems as they save a significant amount of computer memory. Pooling, the process of having the filter skip steps as it passes over the image, is another method commonly used to help this

endeavor [4]. Regularization techniques, those methods used to minimize over-fitting are another important feature of CNNs. The most common forms of regularization are batch normalization, weight decay, and dropout, all of which will be applied in this paper (for further details on these see [4, 7, 10]).

The specific type of CNN that we will use to approach this problem is a residual CNN. This type of network has layers that are connected not only to the next layer, but layers further on as well to aid in the backward propagation of the error. This allows for much deeper networks to be trained more quickly [5]. A residual CNN was one of the first networks to surpass human level performance on the ImageNet database; a database with over a million images and thousands of categories. We will modify this architecture to predict the $6N$ parameters within an IFS consisting of N functions.

4 Fractal Databases

Vast amounts of data are required to train a CNN without over-fitting. As outlined in the previous sections, we created databases of fractals, organized by the number of functions in the IFSs. This was done with our own code written in the C programming language and is available upon request. The functions in each IFS were also ordered based on their parameter values so as to have consistency when computing the error in the output. Specifically, the functions were listed in increasing order of their a_i parameter as depicted in Eq. 1. If two functions were to have the same value for that parameter, a very unlikely case for randomly generated double precision values, then the one with the smallest b_i parameter would be listed first. Within the databases we stored the parameters of the IFS as well as the location of the average pixel, the standard deviation of the fractals' pixels in both the horizontal and vertical directions, the fractal dimension, number of pixels corresponding to the fractal, and the 640×640 binary image. Here, we use the terminology average pixel to signify the centroid location of the fractal in terms of pixel coordinates if each of the pixels corresponding to the fractal possessed a unit mass. Similarly, the standard deviation of the pixels in a given direction is a measure of the spread of those unit mass pixels from the average pixel location.

We can analyze how the parameters of a fractal affect its properties, and how these properties change with varying numbers of functions in the IFSs. We call the values within the matrix of each affine transformation of an IFS the multiplicative parameters. From Fig. 1 we can see that for fractals with fewer functions in their IFS, the magnitude of the multiplicative parameters is related, on average, to the number of pixels corresponding to the fractal. However, as the number of functions in an IFS increases, the affect of the magnitude of the parameters diminishes. We see the same results from Fig. 2 that examines those fractals with greater numbers of pixels corresponding to the fractal. We can also analyze how properties of the fractals change with varying numbers of functions in an IFS. Examining the distributions of the number of pixels corresponding to the fractals as shown in Fig. 3, we see that

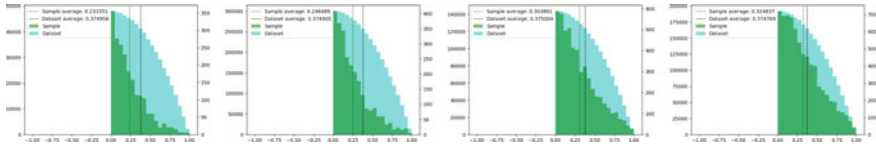


Fig. 1 A comparison of the magnitude of the multiplicative parameters as the number of of functions in an IFS increases. The green represents the distribution of a subset of these parameters that have the fewest number of pixels corresponding to the fractal, whereas the blue represents the distribution of the entire database with that number of functions in an IFS. Going from left to right, the graphs show these distributions for fractal databases with 2, 4, 6, and 8 functions in the IFSs

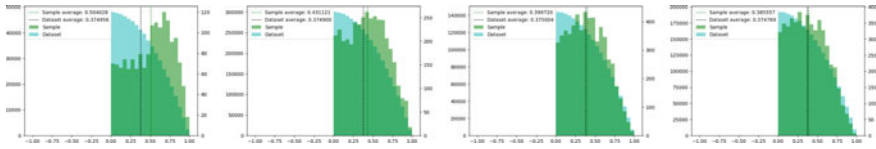


Fig. 2 A comparison of the magnitude of the multiplicative parameters as the number of of functions in an IFS increases. The green represents the distribution of a subset of these parameters that have the largest number of pixels corresponding to the fractal, whereas the blue represents the distribution of the entire database with that number of functions in an IFS. Going from left to right, the graphs show these distributions for fractal databases with 2, 4, 6, and 8 functions in the IFSs

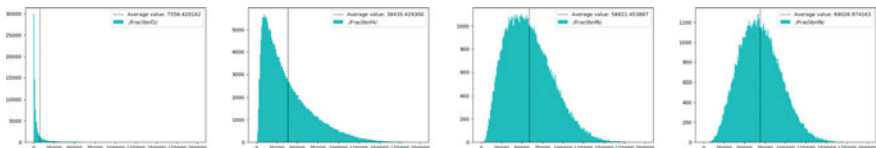


Fig. 3 A comparison of the distributions of the number of pixels corresponding to a fractal as the number of functions in an IFS is changed. From left to right there are 2, 4, 6, and 8 functions in the IFSs

on average, the number of pixels increases with an increasing number of functions in the IFS. Additionally, from Fig. 4 we see that the standard deviation of the pixels decreases, on average, with an increasing number of functions in the IFS. Pairing these concepts together, as the number of functions in an IFS is increased we go from fractals that are spread out with fewer numbers of pixels, to fractals that are more compact with higher numbers of pixels.

Of course, these trends will disappear once the number of functions in the IFSs is increased to a certain threshold. The standard deviation of the pixels corresponding to the fractal will have to increase with the increasing number of pixels, and the increasing number of pixels will have to stop once the entire image begins to be taken up by the fractal.

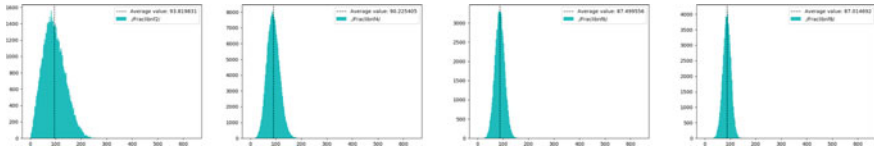


Fig. 4 A comparison of the distributions of the standard deviation of the pixels corresponding to the fractals along the horizontal direction as the number of functions in an IFS is changed. From left to right there are 2, 4, 6, and 8 functions in the IFSs

5 Results

Using the concepts outlined in Sect. 3, many network architectures were tested with the following general structure:

- Take as input the 640×640 binary image.
- Apply the hidden layers (at least 20 were used in our simulations) with 5 layers of pooling
- Apply a fully connected layer leading to the $6N$ outputs
- Calculate the loss using the mean squared error.

Previous studies of this problem often used the Hausdorff distance as a measure of the similarity between two attractors as opposed to the mean squared error of the parameter values. By reconstructing the image from the predicted parameter values, a possible extension of this work could be to include this metric in the loss function. However, this method would be computationally expensive, and there is no guarantee that the model would predict parameters resulting in contractive functions, especially at the beginning of training.

Between each layer of the model, batch-normalization was used to aid with internal covariate shift, as outlined in [7]. Additionally, dropout with a 50% probability was applied to the fully connected layer, and L^2 weight decay was added to the loss function. All neural network components of this paper were implemented using python version 3.7 along with version 1.2 of the pytorch library; this code is available upon request. To measure the accuracy of the models, we used various tolerance levels and deemed the output correct if it was within the given tolerance. That is, if 12 out of 24 of the models predicted parameters were within 0.1 of the true value, we would say it has 50% accuracy at that tolerance level.

Table 1 shows sample output from a network trained on IFSs consisting of four functions with an input it had not yet seen. Visually, the fractals obtained are quite different. However, comparing parameters, more than half are within 0.3 of their respective true value.

From Table 2 we can obtain a more accurate representation of how the models perform overall. There is a strange trend in this table that the models are becoming more accurate with increasing numbers of functions in the IFSs. This is unexpected because the model has to predict more values, and is performing better. However, none of these accuracies are sufficient to consistently produce IFSs with attractors

Table 1 The top row shows a comparison of the true parameter values for the fractal with four functions in its IFS to the model output. The bottom row shows the corresponding fractal images



True IFS	Model Output
$w_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -0.4193 & -0.0270 \\ -0.4613 & -0.2053 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0.2791 \\ 0.0481 \end{pmatrix}$	$w_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -0.5886 & -0.0300 \\ -0.1213 & -0.0009 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -0.0672 \\ 0.0688 \end{pmatrix}$
$w_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -0.3939 & 0.4496 \\ 0.4406 & 0.4919 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -0.2792 \\ -0.2756 \end{pmatrix}$	$w_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -0.3463 & 0.0072 \\ 0.0261 & -0.0587 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -0.0345 \\ 0.0054 \end{pmatrix}$
$w_3 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.1568 & -0.3538 \\ -0.7952 & 0.3318 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0.4873 \\ 0.2117 \end{pmatrix}$	$w_3 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.0156 & -0.0798 \\ -0.0697 & -0.1472 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0.0220 \\ -0.0234 \end{pmatrix}$
$w_4 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.3812 & 0.5316 \\ -0.5675 & 0.2246 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -0.1638 \\ 0.2334 \end{pmatrix}$	$w_4 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.2440 & -0.1563 \\ -0.0041 & -0.0303 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -0.1560 \\ -0.0047 \end{pmatrix}$
	

Table 2 Accuracy at various tolerance levels of the same model being trained on fractal databases with different numbers of functions in the IFSs. Each model was trained for the same amount of time and accuracy was averaged over 10,000 images the model had not trained on

Tolerance	Accuracy			
	2 functions (%)	4 functions (%)	6 functions (%)	8 functions (%)
0.1	20.93	24.36	25.81	26.56
0.2	40.24	46.09	48.67	50.42
0.3	56.94	62.78	65.25	66.99
0.4	70.66	75.01	76.09	77.08
0.5	81.11	83.54	83.78	84.10

visually similar to the input image. Despite this, neural networks provide results quickly, and the probability of randomly producing an 8-function IFS with at least 84% of its values within 0.5 of the true values is quite small. Hence the neural network outputs could be used to create an initial population for a genetic algorithm or swarm intelligence method.

6 Conclusions and Future Work

The field of neural networks has grown exponentially in recent years. Convolutional neural networks have become more complex and are able to solve many image related problems. Following this trend we trained several neural networks to predict the parameters used to construct fractals made from iterated function systems. In order to do so, we constructed fractal databases and analyzed them to determine features of iterated function system fractals in general. The trained neural networks, while giving outputs resulting in visually different fractals, provided parameters which could be used to initialize other search algorithms. With increasing numbers of functions in the iterated function system, the models performed better. We plan to exploit this trend and create more databases with higher numbers of functions, as well as test other model architectures.

Acknowledgements This research was enabled in part by support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

References

1. Abiko, T., Kawamata, M.: IFS coding of non-homogeneous fractal images using Gröbner basis. In: Proceedings of the IEEE International Conference on Image Processing, pp. 25–29 (1999)
2. Barnsley, M.: Fractals Everywhere, 2nd edn. Academic Press, San Diego (1993)
3. Berkner, K.: A wavelet-based solution to the inverse problem for fractal interpolation functions. In: Lévy Véhel, J., Lutton, E., Tricot, C. (eds.) Fractals in Engineering, pp. 81–92. Springer, London (1997). https://doi.org/10.1007/978-1-4471-0995-2_7
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). <http://www.deeplearningbook.org>
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV (2016). <https://doi.org/10.1109/CVPR.2016.90>
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034. Santiago (2015). <https://doi.org/10.1109/ICCV.2015.123>
7. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015). arXiv preprint [arXiv:1502.03167v3](https://arxiv.org/abs/1502.03167v3)
8. Nettleton, D.J., Garigliano, R.: Evolutionary algorithms and a fractal inverse problem. Biosystems **33**, 221–231 (1994). [https://doi.org/10.1016/0303-2647\(94\)90007-8](https://doi.org/10.1016/0303-2647(94)90007-8)

9. Quirce, J., Iglesias, A., Gálvez, A.: Cuckoo search algorithm approach for the IFS Inverse Problem of 2D binary fractal images. In: Tan, Y., Takagi, H., Shi, Y. (eds.) *Advances in Swarm Intelligence*, pp. 543–551. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61824-1_59
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The J. Mach. Learn. Res.* 1929–1958 (2014)
11. Vrscay, E.R.: Moment and collage methods for the inverse problem of fractal construction with iterated function systems. In: Peitgen, H.O., et al. (eds.) *Fractals Fundam. Appl. Sci.* Elsevier, Amsterdam (1991)
12. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995. Honolulu, HI (2017). <https://doi.org/10.1109/CVPR.2017.634>

Solving Parameter Identification Problems using the Collage Distance and Entropy



Herb Kunze and Davide La Torre

Abstract In this paper, we extend the previous method for solving inverse problems for PDEs using the Generalized Collage Theorem by searching for a set of coefficients that not only minimizes the collage error but also maximizes the entropy. In this extended formulation, the parameter estimation minimization problem can be understood as a multi-criteria problem, with two different and conflicting criteria, the generalized collage error and entropy associated with the unknown parameters. We use the typical approach of scalarization to reduce the multi-criteria program to a single-criteria program by combining all objective functions with different trade-off weights. Numerical examples confirm that the collage method produces good, but sub-optimal, results, and that adding a relatively low-weighted entropy term helps us obtain a better approximation.

Keywords Inverse problems · Collage theorem · Multi-criteria optimization · Entropy

1 Introduction

Recent work has established the Generalized Collage Theorem as a tool for solving inverse problems for variational equations, such as those arising from the weak formulation of PDEs [3, 4]. Despite the fact that it uses completely different mathematical machinery, the theorem received its name because of the strong philosophical connection to similar work for inverse problems for ODEs, using the Collage Theorem, a simple consequence of Banach's fixed point theorem. Across these different settings, many results on collage-based methods have been established [6], including that they produce very good results: in a typical inverse problem, the estimated

H. Kunze (✉)

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

D. La Torre

SKEMA Business School, Lille, France

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_16

parameter values produce a solution that lies close to the target data. On the other hand, it has also been shown that, in general, the estimated parameter values are sub-optimal [1, 2]. In this paper, following on earlier work for the ODEs inverse problem using the Collage Theorem (in fractal imaging [7], and for ODEs inverse problems [5]), we consider adjusting the solution approach for a variational parameter estimation inverse problem by seeking not just to minimize the generalized collage distance with respect to the unknown parameters, but to also maximize the entropy associated with the unknown parameters.

In Sect. 2, we discuss the details of the Generalized Collage Theorem, and in Sect. 3, we explain how we construct the entropy term. Section 4 briefly discusses how we scalarize the two-objective optimization problem. Finally, Sect. 5 presents a numerical example that shows that adding a relatively low-weighted entropy term to the collage distance results in parameter values with a corresponding solution that better agrees with the target data. This conclusion is in agreement with past observations in other mathematical settings.

2 Inverse Problems Using the Generalized Collage Theorem

Many inverse problems for PDEs can be written in the following variational form,

$$a(u, v) = \phi(v), \quad v \in H. \quad (1)$$

where $\phi(v)$ and $a(u, v)$ are linear and bilinear maps, respectively, both defined on a Hilbert space H . Let $\langle \cdot \rangle$ denote the inner product in H , $\|u\|^2 = \langle u, u \rangle$ and $d(u, v) = \|u - v\|$, for all $u, v \in H$. The existence and uniqueness of solutions to this kind of equation are provided by the classical Lax-Milgram representation theorem: Let H be a Hilbert space and ϕ a bounded linear nonzero functional, i.e., $\phi : H \rightarrow \mathbb{R}$. Also suppose that $a(u, v)$ is a bilinear form on $H \times H$ which satisfies the following conditions:

1. There exists a constant $M > 0$ s.t. $|a(u, v)| \leq M\|u\|\|v\|$ for all $u, v \in H$,
2. There exists a constant $m > 0$ s.t. $|a(u, u)| \geq m\|u\|^2$ for all $u \in H$.

Then there is a unique vector $u^* \in H$ such that $\phi(v) = a(u^*, v)$ for all $v \in H$.

While the Lax-Milgram representation theorem gives conclusions on the existence and uniqueness of a solution to the direct problem for an appropriately casted PDE, one now wonders what can be said about the associated inverse problem. That is, suppose that we have a given Hilbert space H , a “target” element $u \in H$ and a family of bilinear functionals a_λ , such that the hypotheses of the theorem are satisfied for each λ . Then by the Lax-Milgram theorem, there exists a unique vector u_λ such that $\phi(v) = a_\lambda(u_\lambda, v)$ for all $v \in H$. We would like to determine if there exists a value of the parameter λ such that $u_\lambda = u$ or, more realistically, such that $\|u_\lambda - u\|$ is small enough. The Generalized Collage Theorem addresses this question.

Theorem 1 (*Generalized Collage Theorem*) [3] *Suppose that $a_\lambda(u, v) : \mathcal{F} \times H \times H \rightarrow \mathbb{R}$ is a family of bilinear forms for all $\lambda \in \mathcal{F}$ and $\phi : H \rightarrow \mathbb{R}$ is a given bounded linear functional. Suppose that the hypotheses of the Lax-Milgram theorem are satisfied for each $\lambda \in \mathcal{F}$, and let u_λ denote the unique solution of the equation $a_\lambda(u, v) = \phi(v)$ for all $v \in H$ as guaranteed by that theorem. Given a target element $u \in H$ then*

$$\|u - u_\lambda\| \leq \frac{1}{m_\lambda} F(\lambda), \quad (2)$$

where

$$F(\lambda) = \sup_{v \in H, \|v\|=1} |a_\lambda(u, v) - \phi(v)|. \quad (3)$$

In order to ensure that the approximation u_λ is close to a target element $u \in H$, we can, by the Generalized Collage Theorem, try to make the term $F(\lambda)/m_\lambda$ as close to zero as possible. If $\inf_{\lambda \in \mathcal{F}} m_\lambda \geq m > 0$ then the inverse problem can be reduced to the minimization of the function $F(\lambda)$ on the space \mathcal{F} , that is,

$$\min_{\lambda \in \mathcal{F}} F(\lambda). \quad (4)$$

To produce a problem that we can actually solve in general, we finite-dimensionalize the problem in (4). Let $V_n = \langle e_1, e_2, \dots, e_n \rangle$ be the finite dimensional vector space generated by e_i , so that $V_n \subset H$. Given a target $u \in H$, let $\Pi_{V_n} u$ be the projection of u on the space V_n . We approximate the true error minimization problem by the projected problem

$$\min_{u_\lambda \in V_n} \|\Pi_{V_n} u - u_\lambda\|.$$

We can write

$$\begin{aligned} \|\Pi_{V_n} u - u_\lambda\| &\leq \left(\frac{1}{m_\lambda} \right) \sup_{v \in V_n, \|v\|=1} |a_\lambda(\Pi_{V_n} u, v) - \phi(v)| \\ &\leq \frac{1}{m_\lambda} \max_{v = \sum_{i=1}^n \alpha_i e_i \in V_n, \|v\|=1} \left[\sum_{i=1}^n \alpha_i^2 \right] \left[\sum_i |a_\lambda(u, e_i) - \phi(e_i)|^2 \right] \\ &= \frac{M}{m} \left[\sum_i |a_\lambda(u, e_i) - \phi(e_i)|^2 \right] = \frac{M}{m} F_n(\lambda) \end{aligned}$$

where $M = \max_{v \in V_n, \|v\|=1} \sum_{i=1}^n \alpha_i^2$. This means that

$$\inf_{\lambda \in \Lambda} \|\Pi_{V_n} u - u_\lambda\| \leq \frac{M}{m} \inf_{\lambda \in \Lambda} F_n(\lambda). \quad (5)$$

From a practical standpoint, then, we choose a value for n and solve

$$\inf_{\lambda \in \Lambda} F_n(\lambda).$$

In the following sections, we use the abbreviation CD (for “collage distance”) to denote the function $F_n(\lambda)$.

3 Entropy

The classical measure of information is Shannon entropy, taking the form $\sum_i p_i \ln(p_i)$, where $0 < p_i < 1$. We adapt this form to our unknown parameter λ_i through a simple scaling. For a set of coefficients $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$, we defined:

$$ENT(\lambda) = - \sum_1^m \frac{|\lambda_i|}{\Lambda} \ln \frac{|\lambda_i|}{\Lambda} \quad (6)$$

where the constant $\Lambda \geq \sum_i |\lambda_i|$. In practical terms, one can think that the allowable values of λ live inside the m -dimensional hypercube with edge-length 2Λ . In order to maximize this entropy term, we minimize its opposite or negative—also known as “negentropy.”

4 Multicriteria Optimization and Our Scalarized Problem

We recall some basic results from multicriteria optimization [8] In an abstract setting, a finite-dimensional multicriteria optimization problem has the form

$$\max J(x), x \in X \quad (7)$$

where $(X, \|\cdot\|)$ is a Banach space, $J : X \rightarrow \mathbb{R}^p$ is a vector-valued functional, and \mathbb{R}^p is ordered by the Pareto cone \mathbb{R}_+^p . A point $x \in X$ is said to be Pareto optimal or efficient if $J(x)$ is one of the maximal elements of the set of achievable values $J(X)$. Thus a point x is Pareto optimal if it is feasible and, for any possible $y \in X$, $J(x) \leq_{\mathbb{R}_+^p} J(y)$ implies $x = y$. In a more synthetic way, a point $x \in X$ is said to be Pareto optimal if $(J(x) + \mathbb{R}_+^p) \cap J(X) = \{J(x)\}$.

Scalarization is probably the simplest and most widely used technique to identify Pareto optimal solutions. One solves the scalar problem

$$\max \beta \cdot J(x), x \in X \quad (8)$$

where $\beta \in \text{int}(\mathbb{R}_+^p)$. Since the Pareto optimal solution depends on β , by varying β it is possible to obtain different Pareto optimal points. In the case that each component J_i is concave, J is a vector-valued concave functional, and, under this assumption, the scalarized problem (8) is also concave. This means that we can find Pareto optimal points of a concave problem by solving a concave scalar optimization problem, and for each $\beta \in \text{int}(\mathbb{R}_+^p)$ different Pareto optimal points can be obtained. For concave problems, the converse of this result is only partially true, since for every Pareto optimal point \bar{x} , there is a nonzero $\bar{\beta} \in \mathbb{R}_+^p$ such that \bar{x} is a solution of the scalarized problem (8) with $\beta = \bar{\beta}$.

For the purpose of this paper, we consider the scalar problem

$$\min_{\lambda \in \Lambda} \beta_1 CD(\lambda) - \beta_2 ENT(\lambda) \quad (9)$$

5 Numerical Results

Example 1 We consider the boundary value problem (BVP)

$$-\frac{d}{dx} \left(K(x) \frac{du}{dx}(x) \right) = f(x), \quad u(0) = 0, \quad u(1) = 0 \quad (10)$$

with $f(x) = -12x + 1$ and true diffusivity and solution $K_{true}(x) = 1 + 3x$ and $u_{true}(x)$. We sample $u_{true}(x)$ at 10 uniformly distributed points in $[0, 1]$. We add 1% relative noise to these 10 values to produce 10 data points, and we fit a fourth-degree polynomial to the points to produce a target function $u_{target}(x)$. This results of this process are displayed in Fig. 1.

Now, we seek to solve the inverse problem: given $u_{target}(x)$ and $f(x)$, estimate $K(x)$ of the form

$$K(x) = \sum_{i=0}^4 \lambda_i x^i$$

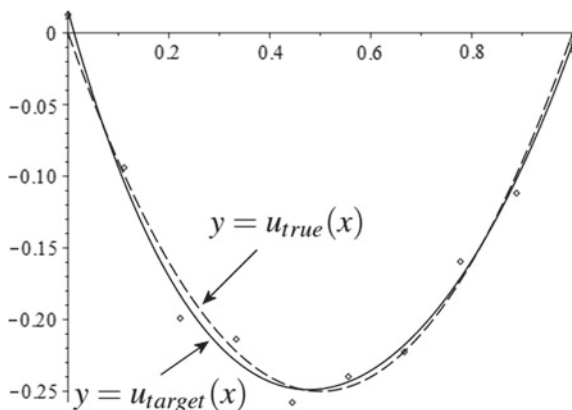
so that $u_{target}(x)$ is an approximate solution to the BVP.

We view the BVP as a steady-state heat equation and recast it its weak formulation, $a(u, v) = \phi(v)$, with

$$a(u, v) = \int_0^1 K(x) u'(x) v'(x) dx, \quad \text{and} \quad (11)$$

$$\phi(v) = \int_0^1 f(x) v(x) dx. \quad (12)$$

Fig. 1 Graphs of $y = u_{true}(x)$ (dashed), $y = u_{target}(x)$ (solid), with the 10 noised data values (diamonds)



where $v, u \in H_0^1([0, 1])$. We project the problem onto

$$V_n^1 = \{v \in C[0, 1] : v \text{ is linear on } [x_{i-1}, x_i], i = 1, \dots, n + 1, v(0) = v(1) = 0\}$$

which has as a basis the so-called hat functions

$$e_i(x) = \begin{cases} (n + 1)(x - x_{i-1}), & x_{i-1} \leq x \leq x_i \\ -(n + 1)(x - x_{i+1}), & x_i \leq x \leq x_{i+1}, i = 1, \dots, n, . \\ 0, & \text{otherwise} \end{cases}$$

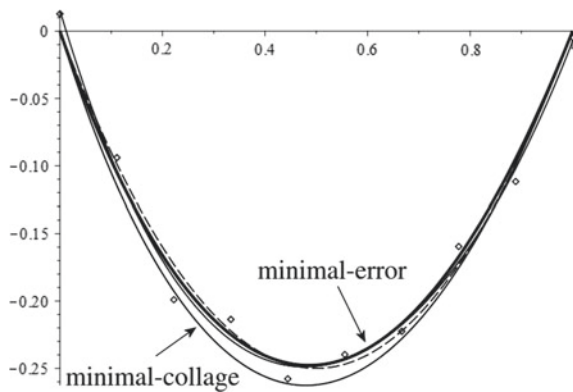
Note that if continuous $K(x) > 0$ for all $x \in [0, 1]$, then m in our formulation can be chosen as $\min_{x \in [0, 1]} K(x)$.

Setting $n = 30$, we use the preceding ingredients to construct $CD(\lambda)$, with $u = u_{target}$, and $ENT(\lambda)$, setting $\Lambda = 25$, and then solve (9) for different choices of β_1 and β_2 . Each solution produces values for λ , and hence $K(x)$, which we use to solve numerically the BVP for u_λ . Finally, we calculate $ER = \|u_{true}(x) - u_\lambda(x)\|_2$. Table 1 shows the results. The first row of the table gives the results when no entropy is used ($\beta_2 = 0$); note the values of the generalized collage distance and the approximation error. In row two of the table, our objective function has a low-weighted entropy term added. The recovered values of λ change, and, sure enough, the generalized collage distance at this new λ value is higher than the minimal value in row one. But we see that the approximation error decreases. The trend continues as we increase the weighting of the entropy term. In this example, the approximation error continues to improve until we hit “35% entropy,” at which point we see the approximation error start to increase from its minimal value in the table. Figure 2 displays the minimal-collage solution and the minimal-error solution from Table 1, corresponding to $(\beta_1, \beta_2) = (0.7, 0.3)$.

Table 1 Results for Example 1. β_i are the weights, CD is the generalized collage distance, ENT is the entropy, and ER is the error in the solution approximation

β_1	β_2	CD	ENT	ER
1.000	0.000	0.0119685	-0.3736584	0.3861749
0.950	0.050	0.0120116	-0.3753041	0.3401305
0.900	0.100	0.0121578	-0.3770917	0.2921707
0.850	0.150	0.0124380	-0.3790424	0.2431742
0.800	0.200	0.0128938	-0.3811815	0.1955942
0.750	0.250	0.0135814	-0.3835405	0.1561011
0.700	0.300	0.0145780	-0.3861584	0.1400275
0.650	0.350	0.0159919	-0.3890844	0.1637707
0.600	0.400	0.0179775	-0.3923817	0.2232289

Fig. 2 For Example 1, in addition to the information in Fig. 1, we add the minimal-collage solution (which has the smallest minimum in the plot) and the minimal-error solution (thickest curve)



Example 2 We again consider Eq. (10), this time with $f(x) = 2 + \cos x$ and true diffusivity $K_{true}(x) = 1 + 5x^2$. We solve numerically for $u_{true}(x)$, and, as in Example 1, sample the solution at 10 uniformly distribute points in $[0, 1]$, add $\epsilon\%$ relative noise, and fit a fourth-degree polynomial to these data values to produce $u_{target}(x)$. We solve the inverse problem via the minimization problem in (5), as in Example 1. The results for $\epsilon = 0$ are in Table 2, for $\epsilon = 1$ are in Table 3, and for $\epsilon = 3$ are in Table 4. We see similar results to Example 1, although the location in $\beta_1\beta_2$ -space of the minimum error changes a little for the different noise levels.

Table 2 Results for Example 2, with no noise added. β_i are the weights, CD is the generalized collage distance, ENT is the entropy, and ER is the error in the solution approximation

β_1	β_2	CD	ENT	ER
1.000	0.000	0.0021742	-0.4465069	0.1137800
0.975	0.025	0.0022016	-0.4486607	0.0803865
0.950	0.050	0.0022864	-0.4508330	0.0517537
0.925	0.075	0.0024329	-0.4530269	0.0405989
0.900	0.100	0.0026458	-0.4552454	0.0586154
0.875	0.125	0.0029308	-0.4574919	0.0900950
0.850	0.150	0.0032943	-0.4597698	0.1256223
0.825	0.175	0.0037434	-0.4620827	0.1628787
0.800	0.200	0.0042865	-0.4644348	0.2011943

Table 3 Results for Example 2, with 1% relative noise added. β_i are the weights, CD is the generalized collage distance, ENT is the entropy, and ER is the error in the solution approximation

β_1	β_2	CD	ENT	ER
1.000	0.000	0.0021017	-0.4470260	0.0845346
0.975	0.025	0.0021298	-0.4492358	0.0499861
0.950	0.050	0.0022168	-0.4514640	0.0242196
0.925	0.075	0.0023669	-0.4537136	0.0382541
0.900	0.100	0.0025853	-0.4559880	0.0720679
0.875	0.125	0.0028774	-0.4582904	0.1088788
0.850	0.150	0.0032497	-0.4606244	0.1467492
0.825	0.175	0.0037098	-0.4629939	0.1853078
0.800	0.200	0.0042661	-0.4654028	0.2244618

Table 4 Results for Example 2, with 3% relative noise added. β_i are the weights, CD is the generalized collage distance, ENT is the entropy, and ER is the error in the solution approximation

β_1	β_2	CD	ENT	ER
1.000	0.000	0.0023244	-0.4477071	0.0373642
0.975	0.025	0.0023541	-0.4500420	0.0133447
0.950	0.050	0.0024460	-0.4523946	0.0421420
0.925	0.075	0.0026044	-0.4547684	0.0790124
0.900	0.100	0.0028346	-0.4571666	0.1168087
0.875	0.125	0.0031425	-0.4595929	0.1551047
0.850	0.150	0.0035346	-0.4620510	0.1938376
0.825	0.175	0.0040189	-0.4645450	0.2330136
0.800	0.200	0.0046040	-0.4670791	0.2726602

6 Conclusions

In this paper, we consider parameter identification inverse problems for variational equations using the Generalized Collage Theorem (subject to the coercivity constants being bounded away from zero). We give a reminder that making the generalized collage distance very small produces a variational equation with solution close to a given target solution. We incorporate an entropy term to produce a two-criteria optimization problem: the goal is to find parameters that minimize the generalized collage distance and maximize entropy. The point is to demonstrate that the revised problem can yield parameter values that produce a solution that lies even closer to the target solution. We choose to scalarize the two-criteria problem, which leads to a weighted combination of the two criteria as the single objective function. Numerical examples show that a small positive weight on the entropy term indeed produces a better estimate than the case when the weight is zero.

It may bear mentioning in this conclusion that there are other ways to treat such multicriteria optimization problems. One other approach is called goal programming; in this framework, one seeks an acceptable solution (which is likely not optimal) by introducing goals for each of the criteria. Another approach is called the ε -constraint method, which singles out one of the criteria as the sole objective function, while shifting all of the other criteria to equality constraints. The inequalities require each constraint term to be bounded by some small term, which, as the name of the approach suggests, is often denoted ε_i for constraint i .

Acknowledgements Research partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Kunze, H., Vrscay, E.R.: Solving inverse problems for ordinary differential equations using the Picard contraction mapping. *Inverse Problems* **15** (1999)
2. Kunze, H., Hicken, J., Vrscay, E.R.: Inverse problems for ODEs using contraction maps: sub-optimality of the “Collage Method”. *Inverse Problems* **20** (2004)
3. Kunze, H., La Torre, D., Vrscay, E.R.: A generalized collage method based upon the Lax-Milgram functional for solving boundary value inverse problems. *Nonlinear Anal.* **71**, 12 (2009)
4. Kunze, H., La Torre, D., Vrscay, E.R.: Solving inverse problems for variational equations using the “generalized collage methods,” with applications to boundary value problems. *Nonlinear Anal. Real World Appl.* **11**, 5 (2010)
5. Kunze, H., La Torre, D., Vrscay, E.R.: Solving inverse problems for DEs using the collage theorem and entropy maximization. *Appl. Math. Lett.* **25** (2012)
6. Kunze, H., La Torre, D., Mendivil, F., Vrscay, E.R.: *Fractal-Based Methods in Analysis*. Springer (2012)
7. Kunze, H., La Torre, D., Lin, J.: IFSM fractal image compression with entropy and sparsity constraints: a sequential quadratic programming approach. In: *AIP Conference Proceedings*, vol. 1798 (2017)
8. Sawaragi, Y., Nakayama, H., Tanino T.: *Theory of Multiobjective Optimization*. Academic Press (1985)

Mean-Square Stability of Stochastic System with Impulse and Unbounded Delay



Mengling Li, Feiqi Deng, and Xinzhi Liu

Abstract This paper investigates mean-square stability of stochastic systems with time-varying parameters, impulses and unbounded delay. Applying the characteristics of stabilizing and destabilizing impulses, stochastic analysis techniques and mathematical deduction, the two conclusions on mean-square stability analysis of the considered systems are obtained.

Keywords Stochastic system · Impulse · Unbounded delay

1 Introduction

Most of the control systems inevitably have some disturbance factors from interior or outside in the actual operation. Stochastic noise [1, 2], impulse [3–5], unbounded delay [6–8] are often encountered and may destroy the stability of systems. In this paper, we mainly aim at this type of systems which possess many complex factors, such as time-vary parameters, noise, impulses and unbounded delay. Sufficient conditions on mean-square stability for the considered systems are given.

There are many literatures focusing on stability of stochastic systems, impulsive systems and systems with unbounded delay. The mean-square stability of stochastic systems are investigated in the references [9, 10] and the authors in the references [11, 12] apply Razumikhin technique to discuss the moment exponential stability of

M. Li (✉)

School of Mathematics and Big Data, Foshan University, Foshan 528000, China

e-mail: menglingli@aliyun.com

F. Deng

School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China

e-mail: aufqdeng@scut.edu.cn

X. Liu

Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada

e-mail: xzliu@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_17

a class of impulsive stochastic functional differential equations. References [13–16] considered the stability and control of impulsive systems. But to our best knowledge, the mean-square stability of stochastic systems with time-varying parameters, impulse and unbounded delay is not investigated up to now.

In this paper, we aim at the mean-square stability analysis for stochastic systems with impulses and unbounded delay. The two situations for impulses including stabilizing and destabilizing impulses are considered and the sufficient conditions for the mean-square stability of stochastic systems are discussed.

Notations: Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq t_0}, P)$ be a complete probability space with a filtration $\{\mathcal{F}_t\}_{t \geq t_0}$ satisfying usual conditions and $|\cdot|$ be Euclidean norm. Let R denote the set of real numbers, R^+ the set of nonnegative real number, R^n and $R^{n \times m}$ the n -dimensional and $n \times m$ -dimensional real spaces, respectively. $A > 0 (A < 0)$ means that the matrix A is a symmetric positive(negative) definite matrix. N^+ represents the set of positive integers. $D^+\varphi$ stands for the Dini derivative of function φ . $C(\mathcal{G}, \mathfrak{H}) = \{\tilde{h} : \mathcal{G} \rightarrow \mathfrak{H} \text{ is continuous}\}$ and denote by $C^{1,2}([t_0, +\infty) \times R^n; R^+)$ the family of positive real-valued functions defined on $[t_0, +\infty) \times R^n$ which are continuously twice differentiable in $x \in R^n$ and once differentiable in $t \in [t_0, +\infty)$. Denote $a \vee b, a \wedge b$ the maximum and minimum value of a, b , respectively. $E(\cdot)$ denote the mathematical expectation. Let $\Psi = \{\psi(t) \in C(R, R^+ \setminus \{0\}) | \psi(t) \leq 1, t \leq t_0; \psi(t) \geq 1, t > t_0\}$.

2 Problem Formulation

Consider the following stochastic time-varying systems with impulses and time-varying unbounded delay

$$\begin{cases} dx(t) = (A(t)x(t) + B(t)x(t - \tau(t)))dt \\ \quad + (C(t)x(t) + D(t)x(t - \tau(t)))dw(t), & t > t_0 \\ x(t_k) = \alpha_k x(t_k^-), & k \in N^+ \\ x(t) = \phi(t), & t \in [t^0, t_0], \end{cases} \tag{1}$$

where $x = (x_1, \dots, x_n)^T \in R^n$; $A(t) = [a_{ij}(t)]$, $B(t) = [B_{ij}(t)]$, $C(t) = [c_{ij}(t)]$, $D(t) = [d_{ij}(t)] \in C([t_0, +\infty), R^{n \times n})$; $\tau(t) \geq 0$ is a continuous time delay, $\alpha_k \in R^+$ for any $k \in N^+$, $\phi = (\phi_1, \dots, \phi_n)^T \in C([t^0, t_0], R^n)$; $w(t)$ is a one-dimensional \mathcal{F}_t -adapted Brownian motion defined on the complete probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq t_0}, P)$; and $t^0 = \inf_{t \geq t_0} \{t - \tau(t)\}$.

In this paper, we say that if $\alpha_k \in [-1, 1]$, then the corresponding impulse is stabilizing impulse. And the destabilizing impulse means $\alpha_k \in (-\infty, -1) \cup (1, +\infty)$. The mean-square asymptotic stability is the main research content, so we assume that the impulse times t_k satisfy $t_k \rightarrow +\infty$ and $k \rightarrow +\infty$.

Next, Itô formula will be given. For function $V(t, x) \in C^{1,2}([t_0, +\infty) \times R; R^+)$, when $t \neq t_k, k = 1, 2, \dots$, we have

$$dV(t, x(t)) = \mathcal{L}V(t, x(t))dt + \frac{\partial V(t, x(t))}{\partial x}(C(t)x(t) + D(t)x(t - \tau(t))),$$

where operator \mathcal{L} is regular and its detailed definition can be found in the reference [17].

Definition 1 For any initial time t_0 and initial data ϕ , if $\lim_{t \rightarrow +\infty} E|x(t)|^2 = 0$ holds, then the trivial solution of the systems is said to be mean-square stable.

In particular, if there exist a function $\psi(t) \in \Psi$ and a constant M which relates to the initial time and data, such that $E|x(t)|^2 \leq M\psi^{-1}(t)$ for any $t > t_0$, then the trivial solution is said to be mean-square ψ stable.

For convenience's sake, several time-varying matrices are defined as following,

$$\begin{aligned} \Upsilon(t) &= [v_{ij}(t)] = A(t) + A^T(t) + C^T(t)C(t), \\ \Theta(t) &= [\theta_{ij}(t)] = B(t) + C^T(t)D(t), \\ \Pi(t) &= [\pi_{ij}(t)] = D^T(t)D(t), \end{aligned}$$

$\Upsilon(t), \Theta(t), \Pi(t)$ belong to $C([t_0, +\infty), R^{n \times n})$.

3 Main Results

Firstly, we focus on the stabilizing impulses. Before we give the stability conclusion, a very useful lemma is given.

Lemma 1 *If there is a function $\psi(t) \in \Psi$ such that*

$$b(t) \frac{\psi(t)}{\psi(t - \tau(t))} + a(t) < 0, \quad t \geq t_0, \tag{2}$$

then we have

$$E|x(t)|^2 \leq \tilde{X}(t_0)\chi_k\psi^{-1}(t). \tag{3}$$

for $t \in [t_0, t_k)$.

Where

$$\begin{aligned} a(t) &= \max_{i \in \mathcal{N}} \{v_{ii}(t) + \sum_{j=1, j \neq i}^n |v_{ij}(t)| + \sum_{j=1}^n |\theta_{ij}(t)|\} < 0, \\ b(t) &= \max_{i \in \mathcal{N}} \{ \sum_{j=1}^n |\theta_{ji}(t)| + \sum_{j=1}^n |\pi_{ij}(t)| \}, \quad \mathcal{N} = \{1, \dots, n\}, \end{aligned}$$

$$\begin{aligned}\tilde{X}(t_0) &= \sup_{h \in [t^0, t_0]} E(x^T(h)x(h)), \\ \chi_0 &= 1, \quad \chi_k = \prod_{i=0}^{k-1} (1 \vee \alpha_i^2), \quad k \geq 1.\end{aligned}$$

Proof Let $V(t) = x^T(t)x(t) = \sum_{i=1}^n x_i^2(t)$. For any $k \in N^+$ and $t \in (t_k, t_{k+1})$, let $\Delta t > 0$ be small enough such that $t + \Delta t \in (t_k, t_{k+1})$. Then by Fubini theorem, it can be obtained that

$$EV(t + \Delta t) - EV(t) = \int_t^{t+\Delta t} E\mathcal{L}V(s)ds.$$

Since $E\mathcal{L}V(t)$ is continuous in the interval $t \in (t_k, t_{k+1})$, it follows that

$$D^+EV(t) = E\mathcal{L}V(t), \quad t \in (t_k, t_{k+1}), \quad k \in N^+. \quad (4)$$

Based on the definition of operator \mathcal{L} , we can compute

$$\begin{aligned}\mathcal{L}V(t) &= 2x^T(t)(A(t)x(t) + B(t)x(t - \tau(t))) \\ &\quad + (C(t)x(t) + D(t)x(t - \tau(t)))^T(C(t)x(t) + D(t)x(t - \tau(t))) \\ &= x^T(t)(A(t) + A^T(t) + C^T(t)C(t))x(t) \\ &\quad + 2x^T(t)(B(t) + C^T(t)D(t))x(t - \tau(t)) \\ &\quad + x^T(t - \tau(t))D^T(t)D(t)x(t - \tau(t)) \\ &= x^T(t)\Upsilon(t)x(t) \\ &\quad + 2x^T(t)\Theta(t)x(t - \tau(t)) + x^T(t - \tau(t))\Pi(t)x(t - \tau(t)).\end{aligned} \quad (5)$$

Note that

$$\begin{aligned}x^T(t)\Upsilon(t)x(t) &= v_{11}(t)x_1^2(t) + 2v_{12}(t)x_1(t)x_2(t) + \cdots \\ &\quad + 2v_{1n}(t)x_1(t)x_n(t) + v_{22}(t)x_2^2(t) \\ &\quad + \cdots + 2v_{2n}(t)x_2(t)x_n(t) + \cdots + v_{nn}(t)x_n^2(t) \\ &\leq v_{11}(t)x_1^2(t) + |v_{12}(t)|(x_1^2(t) + x_n^2(t)) + \cdots \\ &\quad + |v_{1n}(t)|(x_1^2(t) + x_n^2(t)) + \cdots \\ &\quad + |v_{2n}(t)|(x_2^2(t) + x_n^2(t)) + \cdots + v_{nn}(t)x_n^2(t) \\ &\leq (v_{11}(t) + \sum_{j=2}^n |v_{1j}(t)|)x_1^2(t) + \cdots \\ &\quad + (v_{nn}(t) + \sum_{j=1}^{n-1} |v_{nj}(t)|)x_n^2(t),\end{aligned} \quad (6)$$

$$\begin{aligned}
2x^T(t)\Theta(t)x(t-\tau(t)) &= 2(\theta_{11}(t)x_1(t)x_1(t-\tau(t)) + \cdots + \theta_{n1}(t)x_n(t) \\
&\quad \times x_1(t-\tau(t)) + \theta_{12}(t)x_1(t)x_2(t-\tau(t)) + \cdots + \theta_{1n}(t) \\
&\quad \times x_1(t)x_n(t-\tau(t)) + \cdots + \theta_{nn}(t)x_n(t)x_n(t-\tau(t))) \\
&\leq |\theta_{11}(t)|(x_1^2(t) + x_1^2(t-\tau(t))) + \cdots + |\theta_{n1}(t)| \\
&\quad \times (x_n^2(t) + x_n^2(t-\tau(t))) + |\theta_{12}(t)|(x_1^2(t) + x_2^2(t-\tau(t))) \\
&\quad + \cdots + |\theta_{nn}(t)|(x_n^2(t) + x_n^2(t-\tau(t))) \\
&= \sum_{i=1}^n |\theta_{1i}(t)|x_1^2(t) + \cdots + \sum_{i=1}^n |\theta_{ni}(t)|x_n^2(t) \\
&\quad + \sum_{j=1}^n |\theta_{j1}(t)|x_1^2(t-\tau(t)) + \sum_{j=1}^n |\theta_{jn}(t)|x_n^2(t-\tau(t)),
\end{aligned} \tag{7}$$

$$\begin{aligned}
x^T(t-\tau(t))\Pi(t)x(t-\tau(t)) &= \pi_{11}(t)x_1^2(t-\tau(t)) + 2\pi_{12}(t)x_1(t-\tau(t)) \\
&\quad \times x_2(t-\tau(t)) + \cdots + 2\pi_{1n}(t)x_1(t-\tau(t))x_n(t-\tau(t)) \\
&\quad + \pi_{22}(t)x_2^2(t-\tau(t)) + \cdots + \pi_{nn}(t)x_n^2(t-\tau(t)) \\
&\leq \sum_{j=1}^n |\pi_{1j}(t)|x_1^2(t-\tau(t)) + \cdots + \sum_{j=1}^n |\pi_{nj}(t)|x_n^2(t-\tau(t)).
\end{aligned} \tag{8}$$

Substituting (6), (7), (8) into (5), we can obtain

$$\begin{aligned}
\mathcal{L}V(t) &\leq \max_{i \in \mathcal{N}} \{v_{ii}(t) + \sum_{j=1, j \neq i}^n |v_{ij}(t)| + \sum_{j=1}^n |\theta_{ij}(t)|\} V(t) \\
&\quad + \max_{i \in \mathcal{N}} \left\{ \sum_{j=1}^n |\theta_{ji}(t)| + \sum_{j=1}^n |\pi_{ij}(t)| \right\} V(t-\tau(t)) \\
&= a(t)V(t) + b(t)V(t-\tau(t)),
\end{aligned} \tag{9}$$

for any $k \in N^+$ and $t \in (t_k, t_{k+1})$.

Combining (4) with (9), we have

$$\begin{cases} D^+EV(t) \leq a(t)EV(t) + b(t)EV(t-\tau(t)), & t \neq t_k \\ EV(t_k) = \alpha_k^2 EV(t_k^-), & k \in N^+. \end{cases} \tag{10}$$

Without loss of generality, we assume that $x(t_0) \neq 0$ a.s..

When $t \in [t_0, t_1)$, first it is clear that $EV(t_0) \leq \tilde{X}(t_0)$. If there is $t^* = \inf\{s \in (t_0, t_1) | EV(s) > \tilde{X}(t_0)\chi_1\psi^{-1}(s)\}$, we have

$$EV(s) \leq EV(t^*), \quad \text{if } t_0 \leq s < t^*. \tag{11}$$

$EV(t)$ is derivable when $t \in (t_0, t_1)$, by (11), we can obtain that

$$D^+EV(t^*) = \lim_{\Delta \rightarrow 0^+} \frac{EV(t^* + \Delta) - EV(t^*)}{\Delta} = \lim_{\Delta \rightarrow 0^-} \frac{EV(t^* + \Delta) - EV(t^*)}{\Delta} \geq 0. \tag{12}$$

On the other hand, by (10), (2) and (11),

$$\begin{aligned} D^+EV(t^*) &\leq a(t)EV(t^*) + b(t)EV(t^* - \tau(t^*)) \\ &\leq a(t^*)\tilde{X}(t_0)\chi_1 \frac{1}{\psi(t^*)} + b(t^*)\tilde{X}(t_0)\chi_1 \frac{1}{\psi(t^* - \tau(t^*))} \\ &\leq \tilde{X}(t_0)\chi_1(a(t^*) + b(t^*) \frac{\psi(t^*)}{\psi(t^* - \tau(t^*))}) < 0, \end{aligned} \tag{13}$$

which is contradicted with (12), so we can conclude that (3) holds for $t \in [t_0, t_1)$.

Now, we assume that $EV(t) \leq \tilde{X}(t_0)\chi_{k-1} \frac{1}{\psi(t)}$ holds for $t \in [t_0, t_{k-1})$ and we show that $EV(t) \leq \tilde{X}(t_0)\chi_k \frac{1}{\psi(t)}$ holds for $t \in [t_0, t_k)$.

According to condition and (10), $EV(t_{k-1}) = \alpha_{k-1}^2 EV(t_{k-1}^-) \leq \alpha_{k-1}^2 \tilde{X}(t_0)\chi_{k-1} \times \psi^{-1}(t_{k-1}) \leq \chi_k \tilde{X}(t_0)\psi^{-1}(t_{k-1})$. Then similar to the procedure and method of (11), (12) and (13), we can conclude that $EV(t) \leq \tilde{X}(t_0)\chi_k \psi^{-1}(t)$ holds for $t \in [t_0, t_k)$, $k \in N^+$, which means the conclusion (3).

According to the above lemma, the first theorem is given.

Theorem 1 *If the impulses are all stabilizing or there are only finite destabilizing impulses, then under conditions in Lemma 1, the systems (1) can achieve mean-square ψ stability.*

Proof If the impulses are all stabilizing or there are only finite destabilizing impulse, then there is a positive constant $M \geq \tilde{X}(t_0)\chi_k$ for any $k \in N^+$.

Remark 1 It can be seem from (2) and the structure of $a(t)$, $b(t)$ that the coefficient of non-delay term $A(t)$ plays a critical role for moment stability conclusion. Both the time-delay term and stochastic term have a negative effect on moment stability performance.

Unless otherwise specified, the same sign stands for the same definition. It there are infinite destabilizing impulses, the following lemma is useful.

Lemma 2 *If there are constants $\lambda > 0$, $\beta > 1$ and $\varpi \in (0, \lambda)$ such that*

$$b(t)\beta^{\lambda\tau(t)} + \lambda \ln \beta + a(t) < 0, \tag{14}$$

and

$$|\alpha_k| \leq \beta^{\frac{1}{2}\varpi(t_{k+1}-t_k)}. \tag{15}$$

Then we have

$$\beta^{\lambda(t-t_0)} E(x^T(t)x(t)) \leq \tilde{X}(t_0)\beta^{\varpi(t_k-t_0)}, \quad t \in [t_{k-1}, t_k], k \in N^+. \quad (16)$$

Proof Let $V(t) = x^T(t)x(t)$, $Z(t) = \beta^{\lambda(t-t_0)} EV(t)$, the conclusion (16) is equivalent to prove that the following inequality

$$Z(t) \leq \tilde{X}(t_0)\beta^{\varpi(t_k-t_0)}, \quad (17)$$

holds for $t \in [t_{k-1}, t_k], k \in N^+$.

First, we will prove that (17) holds for $t \in [t_0, t_1)$. It is obvious that $Z(t_0) = EV(t_0) \leq \tilde{X}(t_0)$. If (17) does not hold, then set $t^* = \inf\{s \in (t_0, t_1) | Z(s) > \tilde{X}(t_0)\beta^{\varpi(t_1-t_0)}\}$. Then

$$Z(s) \leq Z(t^*), \quad t^0 \leq s \leq t^*. \quad (18)$$

$Z(t)$ is derivable when $t \in (t_0, t_1)$, so

$$D^+ Z(t^*) = \lim_{\Delta \rightarrow 0^+} \frac{Z(t^* + \Delta) - Z(t^*)}{\Delta} = \lim_{\Delta \rightarrow 0^-} \frac{Z(t^* + \Delta) - Z(t^*)}{\Delta} \geq 0. \quad (19)$$

On the other hand, by (10), (14) and (18), we have

$$\begin{aligned} D^+ Z(t^*) &= \lambda\beta^{\lambda(t^*-t_0)} \ln \beta EV(t^*) + \beta^{\lambda(t^*-t_0)} D^+ EV(t^*) \\ &\leq \lambda\beta^{\lambda(t^*-t_0)} \ln \beta EV(t^*) + \beta^{\lambda(t^*-t_0)} (a(t^*)EV(t^*) + b(t^*)EV(t^* - \tau(t^*))) \\ &= \lambda \ln \beta Z(t^*) + a(t^*)Z(t^*) + \beta^{\lambda\tau(t^*)} b(t^*)Z(t^* - \tau(t^*)) \\ &\leq (\lambda \ln \beta + a(t^*) + \beta^{\lambda\tau(t^*)} b(t^*))Z(t^*) < 0, \end{aligned} \quad (20)$$

which is contradicted with (19). So $Z(t) \leq \tilde{X}(t_0)\beta^{\varpi(t_1-t_0)}$ for $t \in [t_0, t_1)$. Assume that $Z(t) \leq \tilde{X}(t_0)\beta^{\varpi(t_l-t_0)}$ for $t \in [t_{l-1}, t_l), l = 1, \dots, k-1$ and we will show that $Z(t) \leq \tilde{X}(t_0)\beta^{\varpi(t-t_0)}$ for $t \in [t_{k-1}, t_k)$.

$$\begin{aligned} Z(t_{k-1}) &= \beta^{\lambda(t_{k-1}-t_0)} EV(t_{k-1}) \leq \beta^{\lambda(t_{k-1}-t_0)} \alpha_{k-1}^2 EV(t_{k-1}^-) \\ &\leq \alpha_{k-1}^2 Z(t_{k-1}^-) \leq \tilde{X}(t_0)\beta^{\varpi(t_k-t_0)}. \end{aligned}$$

Then similar to the procedure and method of (18), (19) and (20), the conclusion (16) can be obtained.

Theorem 2 Let the maximum impulse time interval $T_{\max} \triangleq \sup_{k \in N^+} \{t_k - t_{k-1}\} > 0$, under the same conditions as Lemma 2, the conclusion (16) is equivalent to

$$E|x(t)|^2 \leq \tilde{\beta} \tilde{X}(t_0)\beta^{-(\lambda-\varpi)(t-t_0)},$$

for any $t \geq t_0$, where $\tilde{\beta} = \beta^{T_{\max}} > 1$, which means that the trivial solution is mean-square stable.

Proof When $t \in [t_{k-1}, t_k)$, applying $\beta^{\varpi(t_k-t_0)} = \beta^{\varpi(t_k-t+t-t_0)} \leq \bar{\beta} \beta^{\varpi(t-t_0)}$.

Remark 2 (I) According to (15), impulse strength and impulse time interval influence each other. The longer the time interval, the impulse strength can be allowed larger, which is also a intuitive thought.

(II) The existing of maximum impulse interval in Theorem 2 means that the impulse strength cannot be too big.

(III) Compare Theorems 1 and 2, it can be found that there is no restriction on the impulse time interval when impulse strength is relatively small, but there is a connection between impulse time interval and impulse strength when the impulse strength is relatively big.

Acknowledgements The authors would like to thank the National Natural Science Foundation of China Under Grants 62003087, 62073081, 12071074, 61873099, Guangdong Young Innovative Talents Project under Grant 2020KQNCX074 for their financial support.

References

1. Li, M., Deng, F.: Almost sure stability with general decay rate of neutral stochastic delayed hybrid systems with Lévy noise. *Nonlinear Anal. Hybrid Syst.* **24**, 171–185 (2017)
2. Haba, Z.: Stabilization of starobinsky-vilenkin stochastic inflation by an environmental noise
3. Liu, X., Wang, Q.: The method of lyapunov functionals and exponential stability of impulsive systems with time delay. *Nonlinear Anal.* **66**(7), 1465–1484 (2007)
4. Hao, X., Zuo, M., Liu, L.: Multiple positive solutions for a system of impulsive integral boundary value problems with sign-changing nonlinearities. *Appl. Math. Lett.* **82**
5. Cheng, P., Wu, Z., Wang, L.: New results on global exponential stability of impulsive functional differential systems with delayed impulses. in: *Abstract and Applied Analysis*, Vol. 2012, Hindawi (2012)
6. Lakshmikantham, V.: Theory of differential equations with unbounded delay. *Theor. Comput. Sci.* **157**(2), 139–159 (1994)
7. Zhang, J.: Absolute stability of a class of neural networks with unbounded delay. *Neural Netw.* **17**(3), 391–397 (2004)
8. Liu, X., Stechliniski, P.: Hybrid stabilization and synchronization of nonlinear systems with unbounded delays. *Appl. Math. Comput.* **280**(C), 140–161 (2016)
9. Briat, C.: Stability analysis and stabilization of stochastic linear impulsive, switched and sampled-data systems under dwell-time constraints. *Automatica* **74**, 279–287 (2016)
10. Shaikhet, L.: Necessary and sufficient conditions of asymptotic mean square stability for stochastic linear difference equations. *Appl. Math. Lett.* **10**(3), 111–115 (1997)
11. Hu, W., Zhu, Q.: Moment exponential stability of stochastic nonlinear delay systems with impulse effects at random times. *Int. J. Robust Nonlinear Control* **29**(12), 3809–3820 (2019)
12. Hu, W., Zhu, Q., Karimi, H.R.: Some improved Razumikhin stability criteria for impulsive stochastic delay differential systems. *IEEE Trans. Autom. Control* **64**(12), 5207–5213 (2019)
13. Li, X., Cao, J.: An impulsive delay inequality involving unbounded time-varying delay and applications. *IEEE Trans. Autom. Control* **62**(7), 3618–3625 (2017)
14. Stamova, I.M., Ilarionov, R., Vaneva, R.: Impulsive control for a class of neural networks with bounded and unbounded delays. *Appl. Math. Comput.* **216**(1), 285–290 (2010)

15. Chang, Y.-K.: Controllability of impulsive functional differential systems with infinite delay in banach spaces. *Chaos, Solitons & Fractals* **33**(5), 1601–1609 (2007)
16. Hu, W., Zhu, Q., Karimi, H.R.: On the p th moment integral input-to-state stability and input-to-state stability criteria for impulsive stochastic functional differential equations. *Int. J. Robust Nonlinear Control* **29**(16), 5609–5620 (2019)
17. Guo, Y., Zhu, Q., Wang, F.: Stability analysis of impulsive stochastic functional differential equations. *Commun. Nonlinear Sci. Numer. Simul.* **82**, 105013 (2020)

BOLD.R: A Software Package to Interface with BOLD Through R



Nishan Mudalige

Abstract DNA barcoding has been established as a reliable system for identifying and classifying species. Data generated from DNA barcoding is continuously cataloged on the Barcode of Life Data System (BOLD) using samples collected from researchers and institutions around the world. Advances in DNA analysis have led to a rapid increase in the amount of data available for researchers to study and modern statistical techniques are consequently playing an increasingly important role in the analysis of such large volumes of data. Existing methods to import data from BOLD into any statistical software can be inconvenient, time-consuming, or provide limited information. One of the most popular software applications for statistical analysis is R and we developed an R packages called BOLD . R which aims to overcome existing barriers currently preventing the ease of access to information on BOLD directly into statistical software. Our package allows users to access public data directly from BOLD into R via the current API maintained by the BOLD system. In this article we discuss the implementation and benefits of BOLD . R for researchers and scientists.

Keywords Bioinformatics · Statistical software · R · DNA barcoding

1 Introduction

DNA barcoding is a revolutionary technique that is on the frontier of modern science. Short genetic markers in the DNA of an organism act as genetic “barcodes” that are used for the classification of species. The procedure of applying DNA barcodes for identification was developed by Hebert et al. and discussed further in [1]. The Barcode of Life Data System (BOLD) was developed at the Centre for Biodiversity Genomics (CBG) to provide a comprehensive platform for storing, analyzing, cataloging and publishing data related to DNA barcode records. Samples are collected by researchers in the field or gathered from museums and sent to the CBG for sequencing and

N. Mudalige (✉)

University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada
e-mail: mudalign@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_18

187

analysis. DNA barcodes are extracted from these samples and all of the data collected is uploaded on to BOLD. The BOLD ID Engine drives the identification of unknown sequences by using an advanced cutting-edge clustering algorithm developed by Hebert and Ratnasingham [2]. The algorithm incorporates all sequences uploaded to BOLD from public and private projects in order to locate the closest match. We note that sequences from private records are never exposed in order to maintain data security. BOLD also provides a clean, modern, web-based interface for accessing data. The powerful integrated environment offered by BOLD and the ability for DNA barcode extraction and data upload to be performed in-house at the CBG has resulted in BOLD becoming a valuable asset for many academic and scientific institutions. Due to the significance of DNA barcoding and the advantages offered by the BOLD system, the number of users as well as the volume of information stored on BOLD has grown at an exponential rate. BOLD version 1.0 launched in 2005 with approximately 58,000 records from which approximately 42,000 DNA barcodes were successfully obtained and these records represented approximately 15,000 species. The current iteration of BOLD is version 4.0, and as of September 2019, BOLD has information on over 10 million records which contain 7.5 million DNA barcode sequences representing information from over 650,000 animal species and species proxies (i.e. BINs).

Researchers involved with genetic barcoding are often interested in the ability to perform statistical analysis on their data. As a result, modern statistical techniques are becoming progressively essential in data analysis, particularly with large volumes of data. Users can conduct a reasonable degree of statistical analysis on BOLD through the web interface, however in-depth analysis of data may require specialized statistical software. Many statistical packages are available for data analysis and one of the most popular is R (<https://www.r-project.org/>).

R is an open-source programming language designed by statisticians and scientists specifically for statistical analysis. R is free and available for download on Microsoft Windows, macOS, and Linux. The base version of R incorporates many functions to perform standard statistical tests, create statistical models, analyze large and small data sets, manipulate data and create graphical plots. Many specialized packages have been developed by the scientific community and contributed to the R project. These packages are typically useful in performing specific tasks or analyzing distinct types of data. There are almost 15,000 packages available on R through the Comprehensive R Archive Network (CRAN) repository as of September, 2019. The ability to access such a vast and diverse collection of packages provides users with the flexibility and convenience to conduct statistical analysis in the R environment. R is also very effective and efficient at managing and storing data. It also has powerful graphical capabilities which allow the user to create clean and professional plots and figures which are suitable for publication. The results obtained from statistical analysis conducted in R is often accepted by the scientific community for publishing in many journals (this includes running simulations).

Although we have discussed the advantages of both BOLD and R, existing methods to retrieve data from BOLD into R (or any statistical package) are inconvenient, time-consuming or return limited information. The typical scenario would consist

of a user logging in to BOLD, performing queries to locate the required data, saving the desired files into a format that can be read by R, such as a comma separated files (.csv) or extensible markup language (XML) file, and finally reading the data into R. We have therefore introduced a more accessible system to provide convenient and direct access to the data stored on BOLD into R. Our solution is an R package called “BOLD.R”. An alpha version of BOLD.R is available for download at <http://boldsystems.org/BOLD.R>.

2 Methods and Implementation

We developed an R package called BOLD.R which allows users to access public data directly from BOLD into R via current APIs maintained by the BOLD system. A user simply needs to load BOLD.R in R and use the functions provided by our package to retrieve public data into R in real time directly through the R interface. BOLD can be used as a point for data storage and validation and R can be used for analysis. The BOLD ID engine will validate, extract and classify genera through DNA barcoding and BOLD.R will complement BOLD by allowing researchers to conduct in-depth statistical analysis of public data stored on BOLD using R.

Data on BOLD that is accessed using BOLD.R is stored within R in a format called a “data frame”. A data frame is a tabular data structure in R which consists of rows and columns. We chose to store the data retrieved from BOLD as a data frame due to the many many advantages offered by the format. Data frames are flexible since each column can store data of a different type (e.g.. strings, numeric, logical etc.); we are able to attach labels to column headers so we can provide a meaningful name to a column; it is relatively easy to modify the contents of a data frame; the user can make changes to the data in a data frame without affecting the original data on BOLD; there are many methods to filter data by the values in a column of a data frame and R provides straightforward and efficient ways to merge data and remove duplicated values in data frames. Once data is retrieved from BOLD into R, the user can apply functions from the substantial library of other packages available to conduct analysis on their data. We provide several examples in Sect. 3.

3 Results and Discussion

The primary motivation for us to develop BOLD.R was to provide a convenient way to transfer public data from BOLD into R. To achieve this, we developed the `get.public` function which allows the user to obtain public data by providing a taxon, ID number, BIN number, project code, institution name, researcher name, geographic location or genetic marker to obtain the desired data from BOLD. A string or a vector of strings can be entered to return a result. The information is

retrieved and stored as a data frame. An example of obtaining several projects at once is illustrated in Fig. 1.

If multiple parameter values are passed into the `get.public` function, the API will perform an inner join on the conditions.

Several other useful functions are integrated into the `BOLD.R` package. One such function is the `summary.bold` function which provides information regarding the number of unique records, number of projects and number of different primers in a data frame. Figure 2 provides an example of the data displayed with this function.

Another useful function is the `nucleotides` function which provides information about the different types of nucleotide alignments within a data frame. Figure 3 provides an example of the output displayed by this function.

We reiterate that R has the ability to create many professional plots. For illustrative purposes we present some plots that were created in R using `BOLD.R` along with other existing packages. The data used to create all of the plots presented in this paper is public data which anyone can easily access from BOLD. Figures 4 and 5 provide examples of popular types of plots for displaying qualitative data. The bar plot and donut plot in these figures can be created using the functions available on the base version of R.

Using additional packages, we can obtain more sophisticated visualizations of data, such as trees and dendrograms. Figures 6 and 7 provide some examples of plots that can be created using the `ape` and `plyr` libraries on public data retrieved through `BOLD.R`. Figure 6 is a tree created using data obtained from a project with a relatively small number of distinct records and Fig. 7 is a dendrogram in a fan layout that was created using data from a project with a relatively large number of distinct data records. The `ape` package is popular among researchers working with phylogenetic

```
> example.df = get.public(container=c("ACAGA, AICC, PRT"))
```

Fig. 1 An example of a command which will retrieve data from several public projects

```
> example.df = get.public(container=c("ACAGA, AICC, PRT, +
AGBOU"))
> summary.bold(example.df)
No. of records           337
No. of unique record codes  4
No. of duplicate records   0
No. of different primers   3
```

Fig. 2 An example of the output obtained from the `summary.bold` function. The data frame in this example consists of 4 different projects which contain a total of 337 unique records with 3 different primers and it does not contain any duplicates

```
> example.df = get.public(container=c("ACAGA, AICC, PRT"))
> nucleotides(example.df)
[1] "COI-5P_nucraw" "COI-LIKE_nucraw" "CYTB_nucraw"
```

Fig. 3 An example of the output obtained from the `nucleotides` function. The data frame in this example consists of 3 different projects which contain a total of 3 different nucleotides

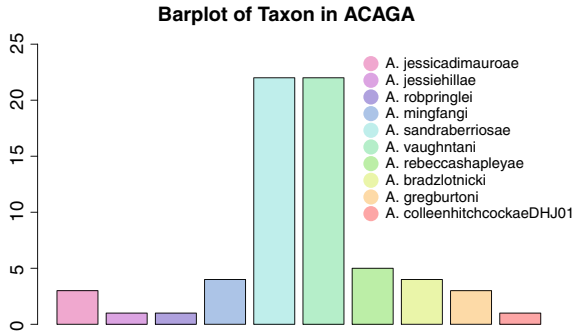


Fig. 4 A bar plot created using records in a data frame. This example contains frequency information about the frequency of specific phyla obtained from a public data set on BOLD

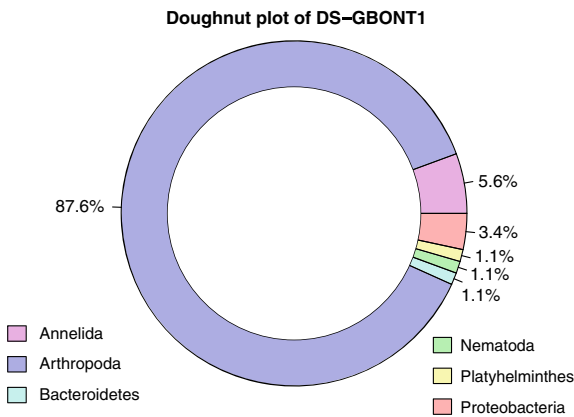


Fig. 5 A doughnut plot created using records in a data frame. This example contains frequency information about the frequency of specific phyla obtained from a public data set on BOLD

data. `ape` stands for Analyses of Phylogenetics and Evolution and it is often used to plot and prune phylogenetic trees, calculate metrics between DNA sequences, read and write nucleotide formats and it also integrates with the BioConductor framework for the analysis of genomic data. More information about `ape` can be found in [4] `plyr` is an R package this is used to efficiently perform split-apply-combine (SAC) procedures on data. It is a very useful package to load into R in order to manipulate very large data sets. More information about `plyr` can be found in [7].

BOLD.R allows the user to create a class of object called a DNAbin by using the `genDNAbin` function which is included with BOLD.R. A DNAbin object can be used with other packages such as `ape` to analyze DNA sequences. Functions in `ape` can be applied on DNAbin objects to create images displaying the alignment of nucleotide sequences, such as the image in Fig. 8. We created Fig. 8 in R using a

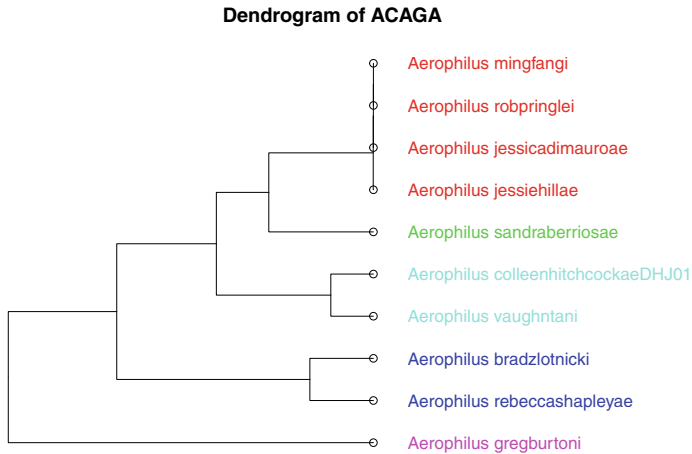


Fig. 6 A tree classifying taxa in a data frame. The additional libraries used to create this plot are `ape` and `plyr`

subset of data retrieved from a public project using `BOLD.R` and then converting the data obtained into a `DNABin` object with the `genDNABin` function.

A few more examples of graphical plots that can be created using `DNABin` objects are given in Figs. 9 and 10. Figure 9 depicts a heatmap and Fig. 10 is a haplotype network plot. The `stringdist` package was used to create the heatmap and `pegas` was used to create the haplotype network. More information about these packages can be found in [6] and [5] respectively.

Samples collected by researchers may also contain geographic information, such as latitude, longitude and elevation. This geographic data can also be plotted in R to provide useful information regarding the specific path that a researcher used to collect samples or to gather information on the spatial distribution of species. The `ggmap` library integrates with the google map API and provides functions which allow users to plots data on maps. More information on `ggmap` can be found in [3]. Various layers of maps can be plotted to show the desired amount of detail. Figure 11 provides visual information about the location in which data was collected by researchers conducting field work in a park for two different projects on `BOLD`. This figure also provides some information about land cover and topography.

`BOLD.R` also provides a function for merging data. The `merge.bold` function efficiently and intelligently merges data frames containing data obtained from `BOLD`. The advantage of using the `merge.bold` function over other functions in R which merge data structures is that `merge.bold` was specifically designed with the structure of the data that returned from `BOLD`. The `merge.bold` function will also remove any duplicate rows that may exist after merging.

We can also create a `FASTA` file directly from data stored on `BOLD` using the `get.fasta` function which is provided with `BOLD.R`. This is a very convenient feature since `FASTA` is a standard text-based format that is used for representing

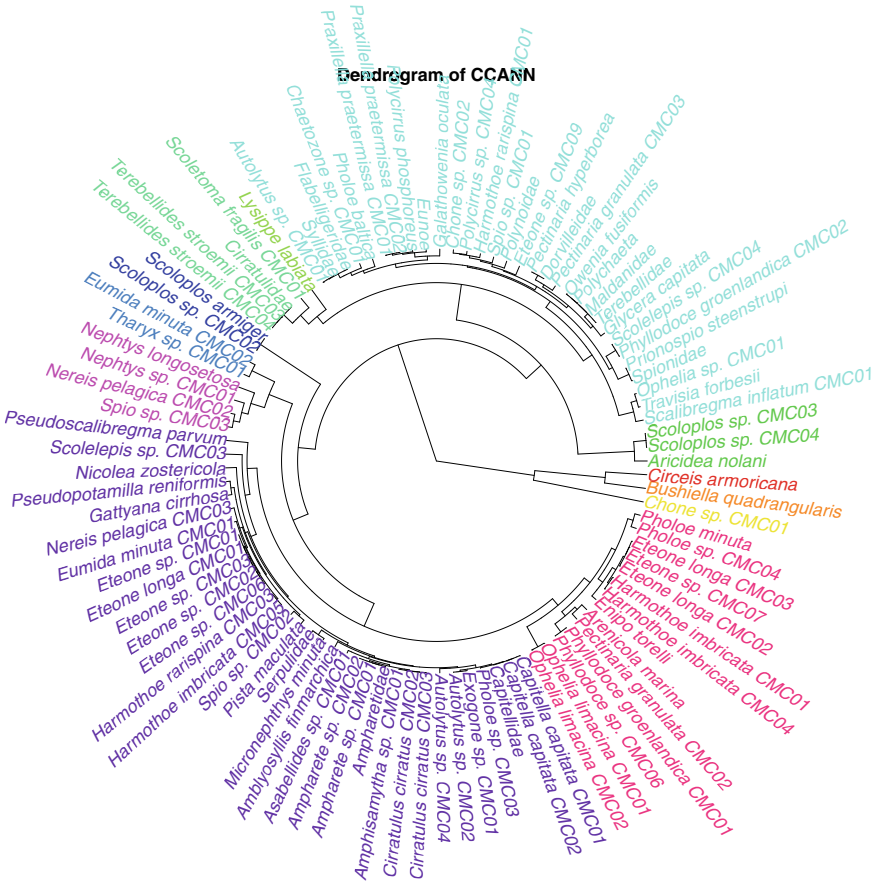


Fig. 7 A dendrogram (fan layout) of the taxa in a data frame. The additional libraries used to create this plot are ape and plyr

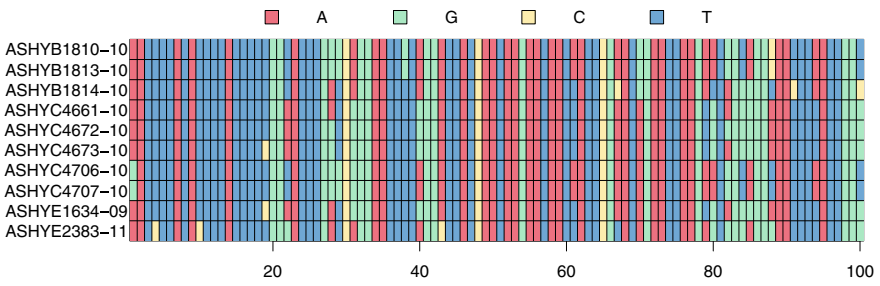


Fig. 8 An image illustrating the alignment of nucleotide created for a small subset of data taken from a public project

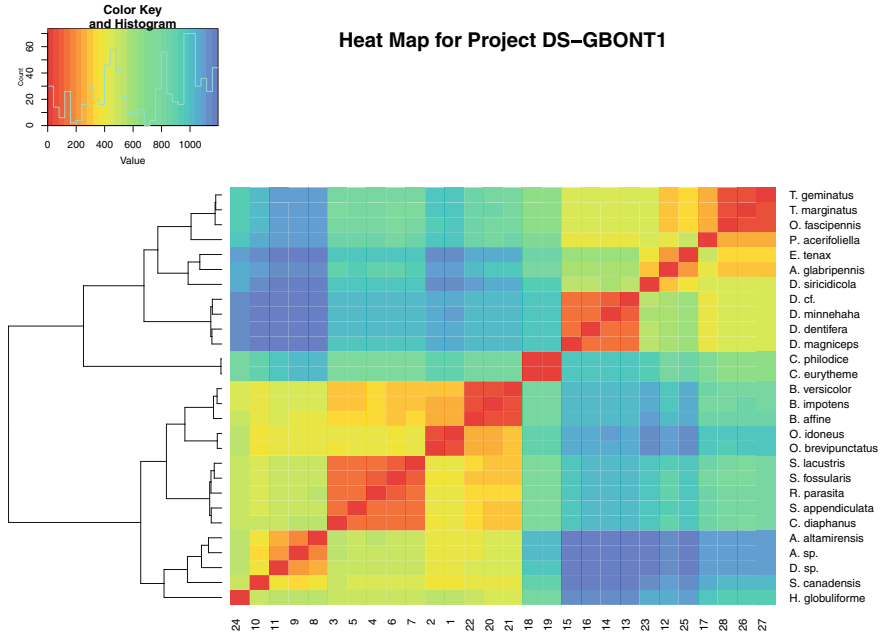


Fig. 9 A heat map for the similarity between barcodes in a public project on BOLD. The additional libraries used to create this plot are `stringdist` and `gplots`

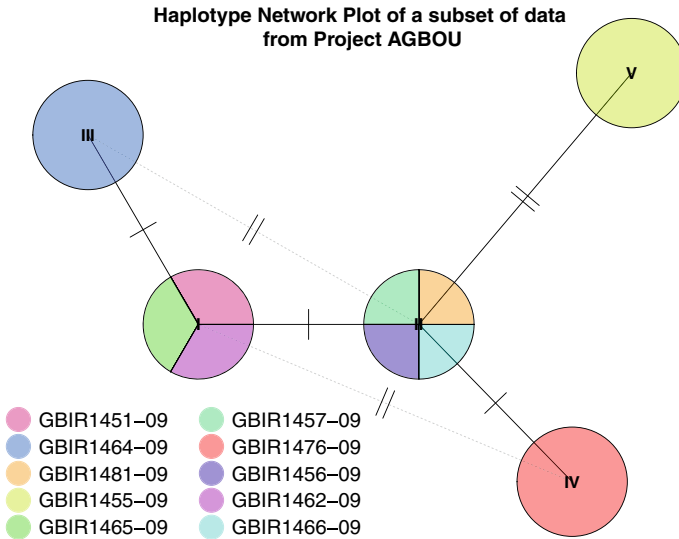


Fig. 10 A haplotype network created using a DNABin created using BOLD.R. The additional libraries used to create this plot are `ape` and `pegas`

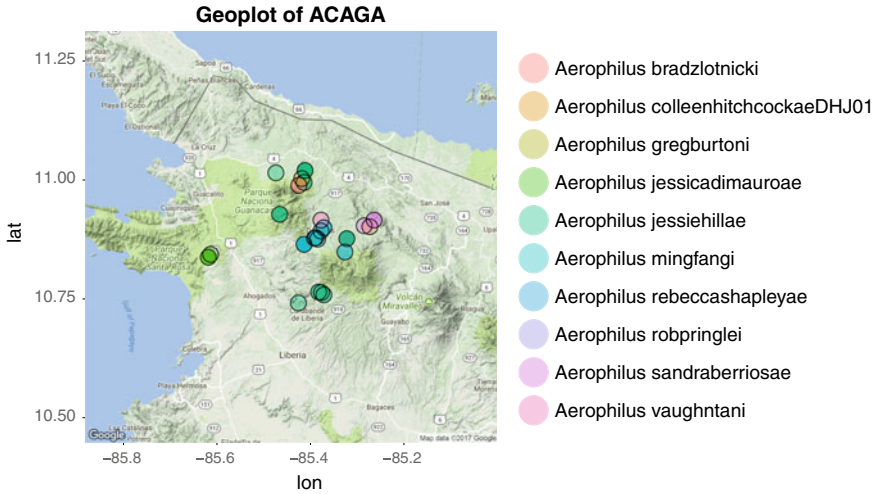


Fig. 11 A geospatial plot indicating locations specimens in a public project stored in BOLD were collected. The additional R library used to create this plot is `ggmap`

nucleotide sequences. The parameters of the `get.fasta` are the same as those in the `get.public` function (i.e. `taxon`, `ID` number, `BIN` number, etc.) and the API will perform an inner join on if multiple parameter values are provided. A single FASTA file can have many records of sequences, and each record consists of the sequence itself as a string of characters and a unique sequence ID. This format allows the storage aligned DNA strings and amino acid sequences making it a simple and flexible file format for bioinformatics.

4 Conclusions

BOLD.R is a useful package which provides the user with the capacity to access and analyze large amounts of public data on BOLD directly through R. The package provides the user with an efficient manner to import and manipulate data on BOLD directly into R, which is the most comprehensive open-source software package available for statistical analysis. BOLD.R can therefore become a valuable tool to assist researchers make informed decisions through data analytics. The package has a diverse array of applications in fields such as genetic barcoding, biodiversity, conservation, population genetics, evolutionary biology, bioinformatics, and education.

Acknowledgements We would like to thank Sujeevan Ratnasingham for his guidance on this project and Megan Milton for providing systematic insight on the structure and operation of BOLD. We would like to thank Ramya Manjunath for her help with testing and developing BOLD.R and for providing assistance understanding the format in which data is stored on BOLD. We would

like to thank Joris D'hondt for his help with compiling the source code into a working binary and Alexandra Stoneham for her help with providing sample project codes for data sets that were used for testing. We also thank Dean Chan and Eddie Ma for their help with setting up the website where the package can be downloaded.

References

1. Hebert, P., Cywinska, A., Ball, S.L., deWaard, J.R.: Biological identifications through DNA barcodes. *Proc. R Soc. Lond.* (2003). <https://doi.org/10.1098/rspb.2002.2218>
2. Hebert, P., Ratnasingham, S.: BOLD: the barcode of life data system. *Mol. Ecol. Resour.* (2007). <https://doi.org/10.1111/j.1471-8286.2006.01678.x>
3. Kahle, D., Wickham, H.: ggmap: spatial visualization with ggplot2. *R. J.* (2013). <https://doi.org/10.32614/RJ-2013-014>
4. Paradis, E., Claude, J., Strimmer, K.: APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* (2004). <https://doi.org/10.1093/bioinformatics/btg412>
5. Paradis, E.: pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* (2010). <https://doi.org/10.1093/bioinformatics/btp696>
6. van der Loo, M.: The stringdist package for approximate string matching. *R. J.* (2014). <https://doi.org/10.32614/RJ-2014-011>
7. Wickham, H.: The split-apply-combine strategy for data analysis. *J. Stat. Softw.* (2019). <https://doi.org/10.18637/jss.v040.i01>

Analysis of Cortical Spreading Depression in Brain with Multiscale Mathematical Models



Hina Shaheen, Roderick Melnik, and Sundeep Singh

Abstract The present study aims at modeling the cortical spreading depression (CSD) propagation in brain considering two different approaches available in the literature: (a) a simplified model consisting of six coupled equations of the reaction-diffusion type in two space dimensions and (b) a one-dimensional, more complex neuronal model comprising of ionic currents and ionic pumps. A study has been conducted to quantify the effects of varying extracellular potassium concentrations on the propagation of CSD in the multiscale reaction-diffusion model by monitoring the respective changes in the extracellular and intracellular concentrations of sodium, chlorine and calcium ions, in addition to evaluating the changes in the generated membrane potentials. In the multiscale neuronal model, the influence of gated conductance on the intracellular and extracellular potassium concentrations of the sodium and potassium and the membrane potential has been reported. The study revealed that the variation in gated conductance results in an increase of the pump currents that leads to the spatio-temporal variations of extracellular potassium.

Keywords Brain · CSD propagation · neuronal model · ion channels · KCl stimulus

H. Shaheen (✉)

MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University (WLU), 75 University Avenue West, Waterloo N2L 3C5, Canada
e-mail: shah8322@mylaurier.ca

R. Melnik · S. Singh

MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University, Waterloo, Canada
e-mail: rmelnik@wlu.ca

S. Singh

e-mail: ssingh@wlu.ca

R. Melnik

BCAM—Basque Center for Applied Mathematics, Bilbao, Spain

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_19

1 Introduction

Cortical spreading depression (CSD), first discovered by Leão [1] in 1944, is a wave that propagates slowly across the cerebral cortex of the brain. This wave can cause a drastic failure of the brain homeostasis leading to temporary impairment in the normal functioning of neurons. The clinical disorders related to CSD can also lead to pathophysiology of various diseases including migraine, ischemic stroke, transient global amnesia, epilepsy and traumatic brain injury [2, 3]. The main property of neural cells is to produce an action potential comprising a rapid increase of the transmembrane potential, called spike, supported by a recovering of the resting condition through a refractory period, where the cell cannot be excited during this period [4]. CSD is a wave of electrophysiological hyperactivity, whereby neurons are first highly excited. That is being followed by a silent phase of membrane hyper-polarization and later the triggered neurons are slowly recovered to their normal frequencies. This neurophysiological phenomenon of CSD results in abrupt changes in the intracellular ion gradients, i.e. an increase in extracellular K^+ and glutamate, along with rise in intracellular Na^+ and Ca^{+2} , followed by sustained depolarization of neurons [5]. Importantly, there are several biophysical, electrophysiological, neurochemical and anatomical elements involved in the propagation of CSD, such as glia, neurons, synapses, cell swelling, many ion channels, ion and transmitter concentrations, pumps, blood vessels, degree of hypoxia, gap junctions, etc. [6].

The mathematical models of CSD have also been explored in the past decades for a better understanding of CSD instigation, propagation and depolarization in the human cortex (e.g.. [7–10] and references therein). In the present studies, two continuum-based multiscale mathematical models of CSD have been analyzed numerically: (a) a simplified two-dimensional model consisting of six coupled equations of the reaction-diffusion type derived from the Tuckwell model [5] in two space dimensions and (b) a one-dimensional multiscale neuronal model comprising of ionic currents and ionic pumps derived from [9]. The mathematical structure of the two-dimensional model considers diffusion in extracellular (ECS) and intracellular (ICS) spaces while the neuronal model considers diffusion in extracellular space only. We demonstrate that the developed models reproduce many important characteristics of CSD over multiple spatio-temporal scales such as instigation, propagation and depolarization of CSD wave. In addition, the effect of change in concentration of high extracellular potassium on the propagation of CSD of the model in two space dimensions has also been evaluated. Finally, one of the main motivations and novelty of this study has been to quantify the effects of varying strength of gated conductances of sodium and potassium on the instigation and propagation of CSD in the one-dimensional neuronal model along with highlighting the effects of higher extracellular potassium in two spatial dimensions.

2 Mathematical Modeling of CSD

In this section the mathematical modeling approaches for the analysis of CSD wave instigation and propagation are discussed in the one- and two-dimensional spaces.

2.1 Reaction-Diffusion Model of CSD in Two-Space Dimensions

A mathematical model is developed for the movements of four basic ions, viz., K^+ , Ca^{+2} , Na^+ , Cl^- , and two neurotransmitter substances, one excitatory (T_E) and the other inhibitory (T_I) in the two-dimensional space (x,y) for simulating the movements of these substance during CSD [5]. Our model assumes that the brain-cell microenvironment can be treated as a porous medium consisting of extra- and intracellular (ECS, ICS) compartments whereby the ions and transmitters are free to diffuse in the extracellular space. The present model is based on reaction-diffusion of ions and neurotransmitters. “Reaction” refers to the ion exchange between ICS and ECS which is the microscopic part of the model at a cellular level and “Diffusion” refers to the ionic propagation between neurons and ECS in the macroscopic part of the model. Thus, the developed model consists of six coupled equations of the reaction-diffusion type which update the extracellular concentrations of ions as follows:

$$\frac{\partial v_i^{ext}}{\partial t} = D_i \nabla^2 v_i^{ext} + F_i(v) \quad i = 1, 2, \dots, 6, \quad (1)$$

where $v_1, v_2, v_3, v_4, v_5, v_6$ are the concentrations of K^+ , Ca^{+2} , Na^+ , Cl^- , T_E and T_I , respectively, at time t , D_i is the diffusion coefficient for the i th component and $F_i(v)$ is the reaction term associated to each ion. Further, it is presumed that the CSD wave results in intense neuronal activity that leads to the abrupt rise in the potassium or calcium ion concentrations of the extracellular compartment. Thus, in the present analysis, the instigation of CSD has been done by specifying the initial condition of K^{ext} as a “supra threshold” Gaussian elevation of potassium chloride concentration with a peak value of 20 mM that is given by:

$$K^{ext}(x, y, 0) = K^{ext,R} + 20 \exp \left[- \left[\left(\frac{x - 1.25}{0.05} \right)^2 + \left(\frac{y - 1.25}{0.05} \right)^2 \right] \right]. \quad (2)$$

We consider two intracellular compartments, one pertaining to synapses and the other pertaining to nonsynaptic processes accounting for contributions from glia. These processes are assigned different ratios of extracellular to intracellular volumes represented by α_1 and α_2 , respectively. Moreover, intracellular ions can only diffuse within a limited region of space or must first become extracellular before becoming free to diffuse over the significant region. The internal ion concentrations (v_i^{int} , $i =$

1, 2, 3, 4) are assumed to be given by the local conservation equations, which for K^+ , Na^+ , Cl^- are (with R being a resting equilibrium):

$$v_i^{int}(x, y, t) = v_i^{int,R} + \alpha_1[v_i^{ext,R} - v_i^{ext}(x, y, t)] \quad i = 1, 3, 4, \quad (3)$$

whereas for Ca^{+2} , we have:

$$v_2^{int}(x, y, t) = v_2^{int,R} + \alpha_2[v_2^{ext,R} - v_2^{ext}(x, y, t)]. \quad (4)$$

The other phenomenological relations for computing the membrane potential, Nernst potential, source and sink terms, pump terms, gated conductance, etc., have been adopted from [5]. In this model, the CSD is initiated based on the potassium hypothesis, whereby the high extracellular potassium concentrations will lead to an increase in the excitability of neurons and promote the further release of potassium. Normally, ion pumps present in the neuron membrane and glia comprehend a set of buffering mechanisms responsible for clearing these extracellular excesses. However, if the concentration exceeds a certain threshold then the process buffers too fast and the resulting mechanism will rise to cope. This reaction-diffusion process depends on both the diffusion of potassium across the extracellular concentrations and the reaction triggered in neighbouring tissue which results in further release of potassium [5]. In what follows, the effect of high extracellular potassium concentration on the propagation of CSD has been quantified utilizing this two-dimensional model ($0 < x < 2.5$ mm, $0 < y < 2.5$ mm) comprising of six coupled reaction-diffusion equation (1).

2.2 One-Dimensional Neuronal Model of CSD

In this section, we will construct a neuronal model that consists of the main characteristics of the more complicated model derived from [9]. The model consists of two ions (sodium and potassium) and two compartments (extracellular and intracellular spaces). According to Kirchhoffs current law [10], the membrane potential V_M (mV) is defined by the ordinary differential equation as:

$$C_m \frac{\partial V_M}{\partial t} = -I, \quad (5)$$

where C_m is the membrane capacitance per unit surface area and I is the total cross membrane ionic current per unit surface area given by the sum of the sodium (I_{Na}), potassium (I_K) and leak currents (I_{Leak}) given by:

$$I_{Na} = I_{Na,T} + I_{Na,P} + I_{Na,Leak} + I_{Na,Pump}; \quad I_K = I_{K,DR} + I_{K,A} + I_{K,Leak} + I_{K,Pump}. \quad (6)$$

In this model, the cross-membrane currents are developed by using Goldman-Hodgkin-Katz (GHK) formulas for the active membrane currents, such as the fast transient sodium current ($I_{Na,T}$), persistent sodium current ($I_{Na,P}$), potassium delayed rectifier current ($I_{K,DR}$) and transient potassium current ($I_{K,A}$). Further, the sodium leak current ($I_{Na,Leak}$), potassium leak current ($I_{K,Leak}$), fixed leak current (I_{Leak}) and sodium-potassium exchange pump currents ($I_{Na,Pump}$ and $I_{K,Pump}$) are computed using the Hodgkin-Huxley (HH) model. The general expressions for the GHK for sodium and potassium currents is given by [11]:

$$I_{ion,GHK} = g_{ion,GHK} m^p h^q \frac{F V_M \left(ion^{int} - \exp\left(\frac{-V_M}{\phi}\right) ion^{ext} \right)}{\phi \left(1 - \exp\left(\frac{-V_M}{\phi}\right) \right)}, \quad (7)$$

where $g_{ion,GHK}$ is the product of the conductance amplitude and membrane permeability for the active currents, m and h are the ion-specific activation and inactivation gating variables and $\phi = RT/F$ is a parameter where R is the universal gas constant, T is the absolute temperature and F is the Faraday's constant. In the HH model, we assumed that the conductances $g_{ion,HH}$, i.e. the conductance amplitude for the passive currents, are constant. Further, the general expression of HH types of currents is given by:

$$I_{ion,HH} = g_{ion,HH} (V_M - V_{ion}), \quad (8)$$

where V_{ion} is the Nernst potential for Na and K ions. The leak currents of all ions in the HH model are summarized in one specific leak current given by:

$$I_{Leak} = g_{HH} (V_M + 70), \quad (9)$$

where g_{HH} is a conductance constant. The rate equations give the dynamic of the gating variables for the potassium activator (m) and sodium inactivator (h) as $dm/dt = \alpha_m(1 - m) - \beta_m m$; $dh/dt = \alpha_h(1 - h) - \beta_h h$. The values of α_i and β_i ($i = m, h$) have been adopted from [9].

The pump currents are given as $I_{Na,Pump} = 3I_{Pump}$ and $I_{K,Pump} = -2I_{Pump}$, where $I_{Pump} = I_{max}/(1 + 2.0[K_{ext}^{-1}]^2(1 + 7.7[Na_{int}^{-1}])^3)$. The model equations in terms of internal and external concentrations coupled with nonlinear diffusion are given as follows:

$$\frac{\partial Na^{ext}}{\partial t} = \frac{A}{F V_{int f}} I_{Na} + D_{Na} \frac{\partial^2 Na^{ext}}{\partial x^2}; \quad \frac{\partial Na^{int}}{\partial t} = -\frac{A}{F V_{int f}} I_{Na}, \quad (10)$$

$$\frac{\partial K^{ext}}{\partial t} = \frac{A}{F V_{int f}} I_K + D_K \frac{\partial^2 K^{ext}}{\partial x^2}; \quad \frac{\partial K^{int}}{\partial t} = -\frac{A}{F V_{int f}} I_K, \quad (11)$$

where A is the cell surface area, V_{int} and V_{ext} are the intracellular and extracellular volumes, respectively, and $f = V_{ext}/V_{int} = 0.15$. The values of other relevant parameters used in this model have been adopted from [5, 9]. In what follows, the effect of gated conductances of sodium and potassium has been evaluated on the CSD wave propagation based on the one-dimensional neuronal model described here.

Throughout this section, CSD was triggered by initiating the Gaussian potassium chloride (KCl) wave, i.e. by changing the initial condition of external potassium as:

$$K^{ext}(x, 0) = K^{ext,R} + K^{max} \exp\left(\frac{-x^2}{2\tau^2}\right), \quad (12)$$

where $\tau = 0.5 \times 10^{-2}$ mm. Motivated by [9], the domain selected for this one-dimensional model was $0 < x < 6$ mm. A finite element method implemented via [12] has been used to solve the set of coupled ordinary and partial differential equations of the multiscale neuronal model and a simplified two-dimensional model.

3 Results and Discussion

In the present paper, a study has been performed for quantifying the effects of different parameters on the CSD wave propagation. In what follows, the effects of sudden changes in KCl stimulus have been quantified on the membrane potential and extracellular concentrations using the model of CSD in two space dimensions. Whereas, the effects of gated conductance on the extracellular potassium concentration have been investigated utilizing the one-dimensional model. This section is split into two parts. Firstly, we will discuss the results obtained with the two-dimensional model, and then with the one-dimensional model, following the description of these models given in the previous section.

3.1 Reaction-Diffusion Model of CSD in Two-Space Dimensions

The temporal response of the extracellular concentrations of Na^+ , K^+ , Ca^{+2} , Cl^{-1} , excitatory and inhibitory transmitters subjected to the initial KCl stimulation applied at the centre of the two-dimensional domain is presented in Fig. 1. As seen from Fig. 1, the KCl stimulus will result in a corresponding increase in the concentration of K^+ and excitatory and inhibitory transmitters, along with a decrease in the Na^+ , Ca^{+2} and Cl^{-1} concentrations. One of our main motivations for the present mathematical study is to evaluate the effects of high extracellular potassium concentration on the propagation of CSD. Primarily, two peak values of K^{ext} in a ‘‘supra threshold’’ Gaussian elevation of KCl concentrations have been selected (a) 40 mM and (b) 60

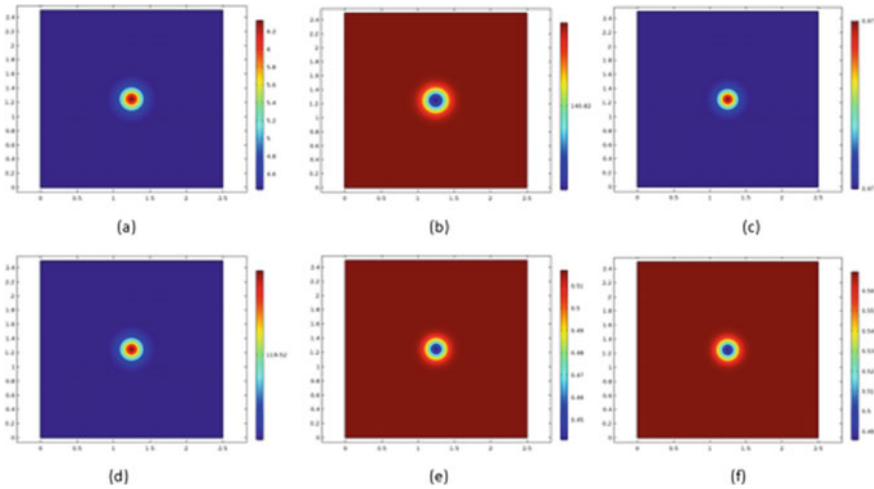


Fig. 1 (Color online) Temporal variation of: **a** Potassium, **b** Chloride, **c** Calcium, **d** Sodium, **e** Inhibitory transmitter and **f** Excitatory transmitter at $t = 2$ s subjected to a *KCl* stimulus

mM. The variation of the peak value of extracellular potassium concentration on the propagation of CSD with time for the two considered cases has been presented in Fig. 2. As evident from Fig. 2, there prevails a significant variation in the predicted membrane potential while the effect on the external potassium concentration is quite negligible. The membrane potential increases as the peak value of the extracellular concentration increases from 20 to 60 mM and vice versa.

3.2 One-Dimensional Neuronal Model of CSD

Here we focus on the analysis of the effects of gated conductance on the extracellular potassium concentration. Based on the model equations (5)–(11) with $x \in (0, 6)$ mm, the spatio-temporal variations in the external potassium concentrations and membrane potential obtained by triggering the CSD with the application of “supra threshold” Gaussian elevation of *KCl* at the left leading edge ($x = 0$), for a peak value of 50 mM at different time steps, are presented in Fig. 3. As evident from Fig. 3, the extracellular potassium concentration increases tremendously due to the instigation of CSD and the depolarization of membrane potential. Importantly, the influence of the CSD is felt more on the left side where the *KCl* stimulation was applied. Similar trends were observed for the membrane potentials as depicted in Fig. 3. The effects of the dynamics gated currents in a neuronal model of CSD by increasing the values of g_{Na} and g_K are presented in Fig. 4. Recall that the variations in the gated currents of the CSD model considered in the present study are the fast inward sodium current and the slower outward potassium current. In a physiological

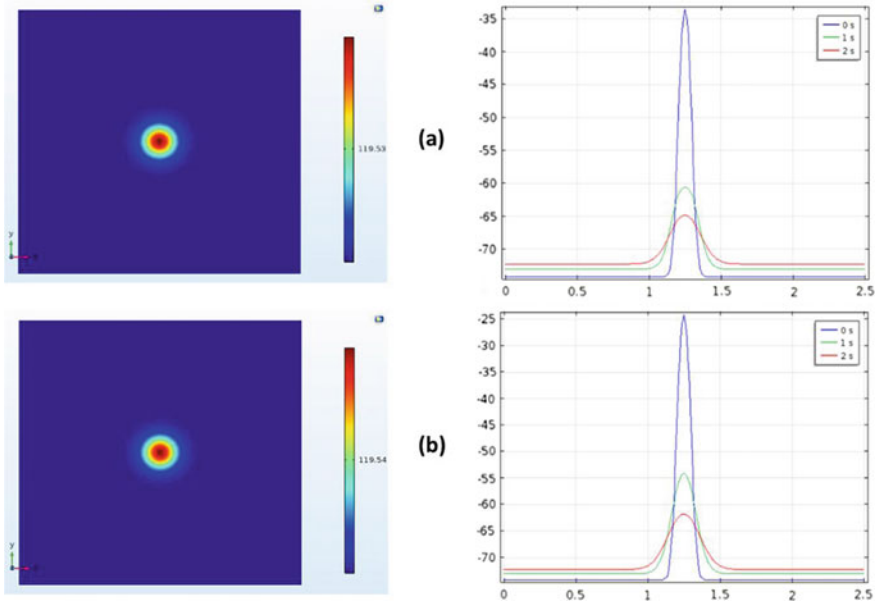


Fig. 2 (Color online) Temporal variation of extracellular potassium concentration (*left*) and membrane potential (*right*) for different values of preset peak value of K^{ext} : **a** 40 mM and **b** 60 mM (on the insert, blue: 0 s, green: 1 s, red: 2 s)

steady state, the gated currents are almost zero and become very strong during the spread of CSD. Even the ionic pump currents are not sufficient to recover the system and alter the CSD to the physiological steady state as presented in Fig. 4. Above all, if the strength of both gated currents is increased by the same factor, a larger pump current will be required to compensate for the gated currents as evident from Fig. 4.

Finally, we note that neural and other electrophysiological activities across the cerebral cortex are stochastic in nature [5], so many sources and sinks of ions and transmitters should be modelled with random processes. A simple way to include these random emissions in the mathematical model of CSD would be by means of a counting process $N(x, t)$, $x_1 \leq x \leq x_2$, $t > 0$ which gives random numbers of action potentials in the time interval $(0, t]$. In this case, the corresponding differential equation for the extracellular potassium ion concentration becomes:

$$\frac{\partial K^{ext}}{\partial t} = \frac{A}{FV_{int}f} I_K + D_K \frac{\partial^2 K^{ext}}{\partial x^2} + \alpha \frac{\partial^2 N(x, t)}{\partial x \partial t}, \quad (13)$$

where α represents the local increase in potassium. Further details in this direction will be reported in our future studies.

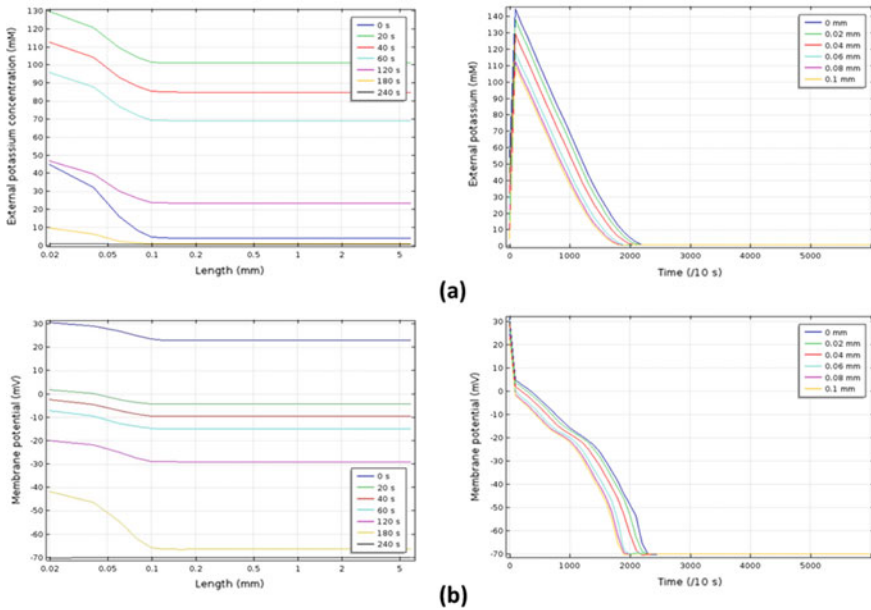


Fig. 3 (Color online) Spatial (*left*) and temporal (*right*) variations of **a** extracellular potassium concentration and **b** membrane potential subjected to initial *KCl* stimulation at $x = 0$ mm (on the insert left, blue: 0 s, green: 20 s, red: 40 s, light blue: 60 s, pink: 120 s, yellow: 180 s, black: 240 s; on the insert right, blue: 0 mm, green: 0.02 mm, red: 0.04 mm, light blue: 0.06 mm, pink: 0.08 mm, yellow: 0.1 mm)

4 Conclusions

In the present study, mathematical models have been developed for quantifying the effects of CSD propagation instigated by sudden changes in the *KCl* stimulus. Two different models have been considered: a multicomponent reaction-diffusion model and a single neuronal model with sodium and potassium currents. We instigated CSD by adding a *KCl* stimulus that leads to an initial condition on extracellular potassium concentration. Adding *KCl* stimulus results in a large increase in the extracellular concentrations of K^+ and a small increase in excitatory and inhibitory transmitters, along with a large decrease in the Cl^- and small decrease in Na^+ , Ca^{+2} extracellular concentrations for the six component model. This behaviour of extracellular concentration of ions is responsible for the propagation of CSD in the brain. The effects of changes in the extracellular potassium concentrations and the gated conductances have also been investigated on the propagation of CSD waves. The results reported in this study could assist in our better understanding of the impact of sudden alterations of the sub-cellular properties on the propagation and instigation of CSD. Future studies will be focused on the inclusion of other compartments, in addition to extracellular and intracellular spaces, such as vascular, glial, neural cell

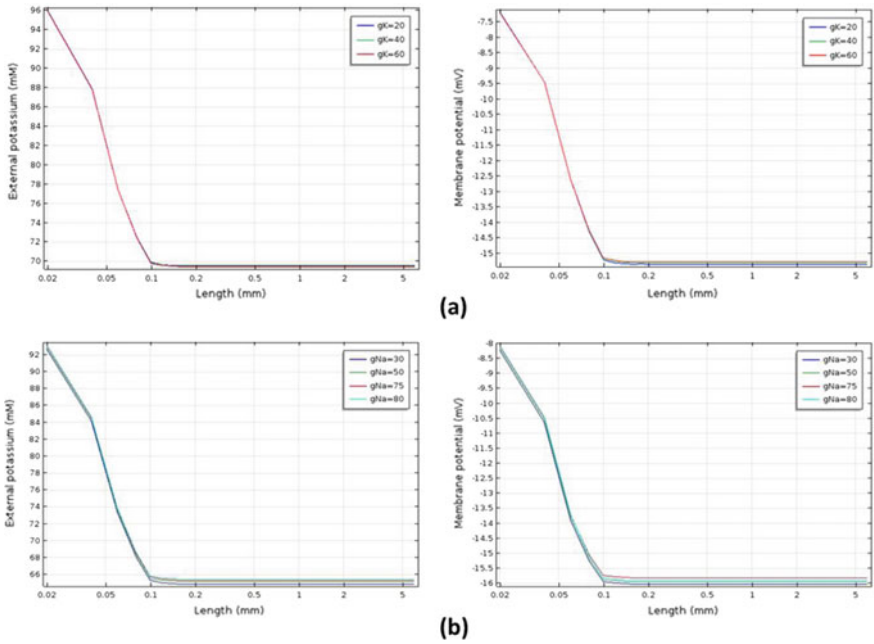


Fig. 4 (Color online) Spatial (*left*) and temporal (*right*) variations of extracellular potassium concentration for different values of potassium and sodium gated conductances: **a** $g_k = 20, 40$ and 60 ; and **b** $g_{Na} = 30, 50, 75$ and 80 (on the insert top, blue: $g_k = 20$, green: $g_k = 40$, red: $g_k = 60$; on the insert bottom, blue: $g_{Na} = 30$, green: $g_{Na} = 50$, red: $g_{Na} = 75$, light blue: $g_{Na} = 80$)

bodies and dendrites, as well as on the development of a new stochastic model based on the ideas highlighted here. This will lead to further clarification of the key mechanisms underlying the dynamics of CSD propagating within the pathological brain.

Acknowledgements Authors are grateful to the NSERC and the CRC Program for their support. RM is also acknowledging support of the BERC 2018-2021 program and Spanish Ministry of Science, Innovation and Universities through the Agencia Estatal de Investigacion (AEI) BCAM Severo Ochoa excellence accreditation SEV-2017-0718.

References

1. Leão, A.A.: Spreading depression of activity in the cerebral cortex. *J. Neurophysiol.* **7**(6), 359–390 (1944)
2. Gerardo, L., Kroos, J.M.: A computational multiscale model of cortical spreading depression propagation. *Comput. Math. with Appl.* **74**(5), 1076–1090 (2017)

3. Cozzolino, O., Marchese, M., Trovato, F., Pracucci, E., Ratto, G.M., Buzzi, M.G., Santorelli, F.M.: Understanding spreading depression from headache to sudden unexpected death. *Front. Neurol.* **9**(19) (2018)
4. Chamanzar, A., George, S., Venkatesh, P., Chamanzar, M., Shutter, L., Elmer, J., Grover, P.: An algorithm for automated, noninvasive detection of cortical spreading depolarizations based on EEG simulations. *IEEE T. Bio-Med. Eng.* **66**(4), 1115–1126 (2018)
5. Tuckwell, H.C.: Stochastic modeling of spreading cortical depression. In: *Stochastic Biomathematical Models*. Springer, Berlin, Heidelberg, pp. 187–200 (2013)
6. Tuckwell, H.C., Hermansen, C.L.: Ion and transmitter movements during spreading cortical depression. *Int. J. Neurosci.* **12**(2), 109–135 (1981)
7. Kager, H., Wadman, W.J., Somjen, G.G.: Conditions for the triggering of spreading depression studied with computer simulations. *J. Neurophysiol.* **88**(5), 2700–2712 (2002)
8. Shapiro, B.E.: Osmotic forces and gap junctions in spreading depression: a computational model. *J. Comput. Neurosci.* **10**(1), 99–120 (2001)
9. Huang, H., Miura, R.M., Yao, W.: A continuum neuronal model for the instigation and propagation of cortical spreading depression. *Bull. Math. Biol.* **73**(11), 2773–2790 (2011)
10. Huang, H., Miura, R.M., Yao, W.: A simplified neuronal model for the instigation and propagation of cortical spreading depression. *Adv. Appl. Math. Mech.* **3**(6), 759–773 (2011)
11. Koch, C., Segev I.: *Methods in Neuronal Modeling from Ions to Networks*. MIT press (1998)
12. COMSOL Multiphysics® v. 5.2. www.comsol.com. COMSOL AB, Stockholm

Impulsive Consensus of Complex-Valued Multi-agent Systems with Hybrid Protocols



Yuan Shen, Xianguo Li, and Xinzhi Liu

Abstract This paper studies the consensus problem of complex-valued multi-agent systems. A complex-valued hybrid consensus protocol with time-delay is proposed, where different network topologies in both continuous-time intervals and impulsive instants are also taken into account. By employing the method of Lyapunov functionals, delay dependent sufficient conditions are established to guarantee that consensus can be achieved in complex domain. Based on various delay sizes of the continuous-time part of the consensus protocol, our result shows that hybrid impulsive protocol leads to consensus if the topological structures of complex-valued multi-agent systems and the impulsive distances can be suitably designed. Numerical simulations are provided to illustrate the effectiveness of the theoretical results.

Keywords Complex-valued multi-agent systems · Consensus · Impulsive protocol · Time delays

1 Introduction

Multi-agent systems have recently been intensively studied in the fields of communication networks, mobile robots, intelligent transportation system, and distributed sensor networks [1–3]. A multi-agent system is a networked system composed of multiple interacting dynamic agents. One of the desired properties in multi-agent systems is consensus among all agents, namely, all agents must reach an agreement

Research supported by NSERC Canada.

Y. Shen (✉) · X. Li · X. Liu
University of Waterloo, N2L 3G1 Waterloo, ON, Canada
e-mail: y57shen@uwaterloo.ca

X. Li
e-mail: xianguo.li@uwaterloo.ca

X. Liu
e-mail: xinzhi.liu@uwaterloo.ca

upon a common value of a certain quantity of interest. Our goal is to design appropriate protocols via distributed coordinated control such that all agents can eventually achieve consensus.

Recently, many protocols have been proposed to solve the consensus problems of multi-agent systems [1, 4–6], and a few of consensus results are derived by designing impulsive consensus protocols based on impulsive control method. In [7], an impulsive consensus protocol is proposed for delay-free linear multi-agent systems with fixed and switching topologies. The hybrid consensus protocol introduced in [8] considers both continuous-time and discrete-time connections in the network topologies, but time-delay is considered only in the impulsive part of the consensus protocol. In [9], a hybrid impulsive protocol is designed, and time delays are taken into account in both continuous-time and discrete-time consensus protocols. Nevertheless, consensus results derived in [9] have no information regarding time-delay in the continuous-time part of hybrid protocols. Moreover, the above consensus results mainly concentrated on multi-agent systems with real variables. Actually, many practical problems in real life can be described more accurately and solved more effectively by complex-valued systems, such as the laser system [10], and the reaction-advection-diffusion system [11]. In particular, it is revealed that the dynamical behaviors for complex-valued neural networks have recently been extensively studied [12–14]. Naturally, there might have potential applications for complex-valued multi-agent systems. Therefore, it is interesting and important to explore the consensus of complex-valued multi-agent systems.

Motivated by the above discussion, a complex-valued hybrid consensus protocol is introduced in this paper. Sufficient conditions are established to guarantee the consensus of complex-valued multi-agent systems with fixed network topologies by employing the Lyapunov method. The constraint of designing impulsive distances based on various delay sizes is also discussed. Numerical simulations are given to illustrate the effectiveness of our theoretical results.

The rest of the paper is organized as follows. In Sect. 2, some background on graph theory and preliminaries for consensus problems and hybrid consensus protocols are presented. In Sect. 3, consensus results for complex-valued multi-agent systems will be established by considering dynamic agents with fixed topology. A typical illustrative numerical example is provided in Sect. 4. Conclusions will be stated in Sect. 5.

Throughout this paper, \mathbb{C}^N denotes the N -dimensional complex vector space. For $x \in \mathbb{C}^N$, the notation x^* refers to the conjugate transpose of x .

2 Preliminaries

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a digraph consisting of N nodes, $\mathcal{V} = \{v_i | i = 1, 2, \dots, N\}$, and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. An edge of \mathcal{G} is denoted by (v_i, v_j) , which means the node v_j can receive information from v_i . The index set $\mathcal{N}_i = \{v_j \in \mathcal{V} | (v_j, v_i) \in \mathcal{E}\}$ denotes the neighbors set of node v_i . A weighted digraph $\mathcal{G}_A = (\mathcal{V}, \mathcal{E}, A)$ is a

digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ associated with a weighted adjacency matrix $A = [\alpha_{ij}] \in \mathbb{R}^{N \times N}$ with nonnegative adjacency elements α_{ij} such that $(v_j, v_i) \in \mathcal{E}$ if and only if $\alpha_{ij} > 0$. It is assumed that $\alpha_{ii} = 0$ for all i . The in-degree and out-degree of node v_i are defined as $d_{in}(v_i) = \sum_{j \in \mathcal{N}_i} \alpha_{ij}$ and $d_o(v_i) = \sum_{j \in \mathcal{N}_i} \alpha_{ji}$, respectively. The graph Laplacian \mathcal{L} of weighted digraph \mathcal{G}_A is defined by $\mathcal{L} = \mathcal{D} - A$, where $\mathcal{D} = \text{diag}\{d_{in}(v_1), d_{in}(v_2), \dots, d_{in}(v_N)\}$. More precisely, $\mathcal{L} = [l_{ij}] \in \mathbb{R}^{N \times N}$, where $l_{ij} = \sum_{j=1}^N \alpha_{ij}$ if $i = j$ and $l_{ij} = -\alpha_{ij}$, otherwise. A weighted digraph \mathcal{G}_A is said to be balanced if $d_{in}(v_i) = d_o(v_i)$ for all i . A digraph is said to be strongly connected if any two distinct nodes of the graph can be connected via a path that respects the direction of the edges of the digraph.

Let $x_i \in \mathbb{C}$ denote the state of node v_i in complex space, and consider each node of a digraph to be a continuous complex-valued agent which has the form

$$\dot{x}_i(t) = u_i(t), \quad i = 1, 2, \dots, N, \quad (1)$$

where $u_i(t) \in \mathbb{C}$ is the control input of agent i . We say that $u_i(t)$ is a protocol if the controller $u_i(t)$ only depends on the state information of node v_i and its neighbors (i.e. $v_j \in \mathcal{N}_i$).

We consider the following complex-valued hybrid impulsive consensus protocol based on fixed digraphs $\mathcal{G}_A = (\mathcal{V}, \mathcal{E}, A)$ at non-impulsive time intervals and $\mathcal{G}'_A = (\mathcal{V}, \mathcal{E}', A')$ at each impulsive instant:

$$u_i(t) = \sum_{j \in \mathcal{N}_i} \alpha_{ij} [x_j(t-r) - x_i(t-r)] + \sum_{k=1}^{\infty} \sum_{j \in \mathcal{N}'_i} \alpha'_{ij} [x_j(t-\bar{\tau}) - x_i(t-\bar{\tau})] \delta(t-t_k), \quad (2)$$

where α_{ij} (α'_{ij}) $\in \mathbb{R}$ is the (i, j) th entry of the weighted adjacent matrix A (A'), and \mathcal{N}_i (\mathcal{N}'_i) denotes the neighbors set of node v_i in digraph \mathcal{G}_A (\mathcal{G}'_A); $\delta(\cdot)$ denotes the Dirac delta function, and impulsive sequence $\{t_k\}$ satisfies $t_0 < t_1 < t_2 < \dots < t_k < t_{k+1} < \dots$, and $\lim_{k \rightarrow \infty} t_k = \infty$; r is the constant time-delay in continuous-time protocol, and $\bar{\tau}$ represents constant impulse delay when processing the impulsive information at each impulsive instant t_k .

Definition 1 We say that a protocol $u_i(t)$ solves the consensus problem if

$$\lim_{t \rightarrow \infty} \|x_i(t) - x_j(t)\| = 0, \quad i, j = 1, 2, \dots, N.$$

By the definition of $\delta(\cdot)$, complex-valued multi-agent system (1) with integrator dynamics under consensus protocol (2) can be described by the following complex-valued impulsive system:

$$\begin{cases} \dot{x}_i(t) = \sum_{j \in \mathcal{N}_i} \alpha_{ij} [x_j(t-r) - x_i(t-r)], & t \neq t_k, \\ \Delta x_i(t_k) = \sum_{j \in \mathcal{N}'_i} \alpha'_{ij} [x_j(t_k - \bar{\tau}) - x_i(t_k - \bar{\tau})], & k \in \mathbb{N}, \\ x_{i_0}(s) = \phi_i(s), & s \in [-\tau, 0], \end{cases} \quad (3)$$

where $\phi_i \in \text{PC}([-\tau, 0], \mathbb{C})$ is the piecewise continuous initial function with $\tau = \max\{r, \bar{\tau}\}$. Without loss of generality, we assume $x_i(t_k) = x_i(t_k^+)$ in the following discussion, which implies solutions of (3) are right continuous at impulsive instants.

If we define state vector $\mathbf{x}(t) = (x_1, x_2, \dots, x_N)^T \in \mathbb{C}^N$, according to interaction topologies $\mathcal{G}_A = (\mathcal{V}, \mathcal{E}, A)$ and $\mathcal{G}'_A = (\mathcal{V}, \mathcal{E}', A')$, system (3) can be rewritten as

$$\begin{cases} \dot{\mathbf{x}}(t) = -\mathcal{L}\mathbf{x}(t-r), & t \neq t_k, \\ \Delta\mathbf{x}(t_k) = -\mathcal{L}'\mathbf{x}(t_k-\bar{\tau}), & k \in \mathbb{N}. \end{cases} \tag{4}$$

Our objective is to derive sufficient conditions on fixed digraphs $\mathcal{G}_A, \mathcal{G}'_A$ and impulsive sequence $\{t_k\}$ to guarantee that complex-valued hybrid protocol (2) can solve the consensus problem.

3 Consensus Results

In order to seek consensus in system (3), we introduce the following disagreement vector $\mathbf{e}(t) \in \mathbb{C}^N$

$$\mathbf{e}(t) = \mathbf{x}(t) - \text{Ave}(x(t)) \cdot \mathbf{1},$$

where $\mathbf{x}(t) = (x_1, x_2, \dots, x_N)^T \in \mathbb{C}^N$, $\mathbf{1}$ denotes the column N-vector with all ones, and $\text{Ave}(x(t)) = \frac{1}{N} \sum_{i=1}^N x_i(t)$. From [9], if \mathcal{G}_A and \mathcal{G}'_A are balanced, we have that $\text{Ave}(x(t))$ is an invariant quantity for $t \geq 0$, say $\text{Ave}(x(t)) = \text{Ave}(x(0)) = \frac{1}{N} \sum_{i=1}^N x_i(0)$. Moreover, the Laplacians \mathcal{L} and \mathcal{L}' have row sum equal to zero, hence $\mathcal{L}\text{Ave}(x(0))\mathbf{1} = \mathcal{L}'\text{Ave}(x(0))\mathbf{1} = \mathbf{0}$.

According to system (4), we can obtain the following impulsive disagreement dynamical system:

$$\begin{cases} \dot{\mathbf{e}}(t) = -\mathcal{L}\mathbf{e}(t-r), & t \neq t_k, \\ \Delta\mathbf{e}(t_k) = -\mathcal{L}'\mathbf{e}(t_k-\bar{\tau}), & k \in \mathbb{N}. \end{cases} \tag{5}$$

For simplicity, we assume all impulses are uniformly distributed (*i.e.* $T = t_k - t_{k-1}$ for all $k \in \mathbb{N}$), and $\bar{\tau} \leq T$ throughout this section.

Theorem 1 *Suppose that \mathcal{G}_A is balanced, \mathcal{G}'_A is balanced and strongly connected. Let $\lambda_2(\mathcal{L}'_s)$ denotes the second smallest eigenvalue of $\mathcal{L}'_s = \frac{\mathcal{L}' + \mathcal{L}'^T}{2}$, if there exist constants $\varepsilon > 0, 0 < \omega \leq 1$ such that*

$$\ln(\alpha + \beta + \omega r) < -cT, \tag{6}$$

where $\alpha = (1 + \varepsilon)(1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2)$, $\beta = (1 + \frac{1}{\varepsilon})(\bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|)^2$, $c = \frac{\|\mathcal{L}'\|^2}{\omega} + \omega$, and $T = t_k - t_{k-1}$. Then, protocol (2) leads to the consensus for agents in complex-valued multi-agent system (1).

Proof Construct the Lyapunov functional candidate $V(t) = V_1(t) + V_2(t)$, where $V_1(t) = \mathbf{e}^*(t)\mathbf{e}(t)$, and $V_2(t) = \omega \int_{t-r}^t \mathbf{e}^*(s)\mathbf{e}(s)ds$ ($0 < \omega \leq 1$). When $t \neq t_k$,

$$\begin{aligned} \dot{V} &\leq 2|\mathbf{e}^*(t)\mathcal{L}\mathbf{e}(t-r)| + \omega\mathbf{e}^*(t)\mathbf{e}(t) - \omega\mathbf{e}^*(t-r)\mathbf{e}(t-r) \\ &\leq 2\|\mathbf{e}(t)\|\|\mathcal{L}\|\|\mathbf{e}(t-r)\| + \omega\mathbf{e}^*(t)\mathbf{e}(t) - \omega\mathbf{e}^*(t-r)\mathbf{e}(t-r) \\ &\leq \left(\frac{\|\mathcal{L}\|^2}{\omega} + \omega\right)\mathbf{e}^*(t)\mathbf{e}(t) \leq \left(\frac{\|\mathcal{L}\|^2}{\omega} + \omega\right)V(t). \end{aligned}$$

Denote $c = \frac{\|\mathcal{L}'\|^2}{\omega} + \omega$, then we can conclude that

$$V(t) \leq V(t_{k-1}) e^{c(t-t_{k-1})}, \quad t \in [t_{k-1}, t_k], \quad k = 1, 2, 3, \dots \quad (7)$$

For $t = t_k$, since we assume that $\bar{\tau} \leq T$, integrate both sides of (5) from $t_k - \bar{\tau}$ to t_k , yields $\mathbf{e}(t_k - \bar{\tau}) = \mathbf{e}(t_k^-) + \int_{t_k - \bar{\tau}}^{t_k} \mathcal{L}\mathbf{e}(t-r)dt$. According to (5), at $t = t_k$, we have that $\mathbf{e}(t_k) = X + Y$, where $X = (I - \mathcal{L}')\mathbf{e}(t_k^-)$, and $Y = -\mathcal{L}\mathcal{L}' \int_{t_k - \bar{\tau}}^{t_k} \mathbf{e}(t-r)dt$, then for any $\varepsilon > 0$, we have the following inequality

$$V_1(t_k) = (X + Y)^*(X + Y) \leq (1 + \varepsilon)X^*X + (1 + \frac{1}{\varepsilon})Y^*Y.$$

Since \mathcal{G}'_A is balanced, strongly connected, and $\mathbf{1}^T \mathbf{e}(t_k^-) = 0$, with similar approaches from [4] and [9], and apply Cauchy Schwarz inequality for integrable complex-valued functions, we can obtain $X^*X \leq (1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2)V_1(t_k^-)$ and $Y^*Y \leq (\bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|)^2 \sup_{s \in [-(\bar{\tau}+r), 0]} V_1(t_k^- + s)$. If we denote $\alpha = (1 + \varepsilon)(1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2)$ and $\beta = (1 + \frac{1}{\varepsilon})(\bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|)^2$, then we have

$$V_1(t_k) \leq \alpha V(t_k^-) + \beta \sup_{s \in [-(\bar{\tau}+r), 0]} V(t_k^- + s), \quad (8)$$

by the continuity of $V_2(t)$,

$$V_2(t_k) \leq \omega r \sup_{s \in [-r, 0]} V(t_k^- + s). \quad (9)$$

If there exists $0 < \omega \leq 1$ such that condition (6) is satisfied, then by IVT, there exists unique $\lambda > 0$ that solves for the following equation

$$\ln[\alpha + \beta e^{\lambda(\bar{\tau}+r)} + \omega r e^{\lambda r}] = -(\lambda + c)T. \quad (10)$$

Since $\lim_{k \rightarrow \infty} t_k = \infty$, there exists integer $p \geq 1$ such that $t_p - \bar{\tau} - r \geq t_0$. For $t \in [t_0, t_p)$:

$$V(t) = V(t)e^{\lambda(t-t_0)}e^{-\lambda(t-t_0)} \leq Me^{-\lambda(t-t_0)}, \tag{11}$$

where $M = \sup_{t \in [t_0, t_p]} V(t)e^{\lambda(t-t_0)}$. Next, we will use the method of mathematical induction to prove that

$$V(t) \leq Me^{-(\lambda+c)(t_{k+1}-t_0)}e^{c(t-t_0)}, \quad t \in [t_k, t_{k+1}), \quad k \geq p. \tag{12}$$

Suppose (12) is true for $p \leq k \leq j$, then

$$V(t) \leq Me^{-(\lambda+c)(t_{k+1}-t_0)}e^{c(t-t_0)}, \quad t \in [t_k, t_{k+1}), \quad p \leq k \leq j. \tag{13}$$

At $t = t_{j+1}$, We will estimate the supremum of $V(t_{j+1}^- + s)$ for $s \in [-(\bar{\tau} + r), 0]$ by considering the following two cases:

Case 1: If $t_{j+1} + s \in [t_0, t_p)$ for some $s \in [-(\bar{\tau} + r), 0]$, then from (11):

$$V(t_{j+1}^- + s) \leq e^{-\lambda s} Me^{-\lambda(t_{j+1}-t_0)} \leq e^{\lambda(\bar{\tau}+\bar{r})} Me^{-\lambda(t_{j+1}-t_0)}.$$

Case 2: If $t_{j+1} + s \geq t_p$ for some $s \in [-(\bar{\tau} + r), 0]$, there exists $\hat{k} (p \leq \hat{k} \leq j)$ such that $t_{j+1} + s \in [t_{\hat{k}}, t_{\hat{k}+1})$, then according to (13),

$$\begin{aligned} V(t_{j+1}^- + s) &\leq Me^{-(\lambda+c)(t_{\hat{k}+1}-t_0)}e^{c(t_{j+1}+s-t_0)} \\ &\leq Me^{-\lambda(t_{j+1}+s-t_0)} \\ &\leq e^{\lambda(\bar{\tau}+r)} Me^{-\lambda(t_{j+1}-t_0)}. \end{aligned}$$

Therefore, we can conclude that $V(t_{j+1}^- + s) \leq e^{\lambda(\bar{\tau}+r)} Me^{-\lambda(t_{j+1}-t_0)}$ for all $s \in [-(\bar{\tau} + r), 0]$, which implies that $\sup_{s \in [-(\bar{\tau}+r), 0]} V(t_{j+1}^- + s) \leq Me^{\lambda(\bar{\tau}+r)} e^{-\lambda(t_{j+1}-t_0)}$. According to (8) and (13), we have

$$V_1(t_{j+1}) \leq [\alpha + \beta e^{\lambda(\bar{\tau}+r)}] Me^{-\lambda(t_{j+1}-t_0)}.$$

Similarly, from (9), (11) and (13), we can show that

$$V_2(t_{j+1}) \leq \omega r e^{\lambda r} Me^{-\lambda(t_{j+1}-t_0)}.$$

Therefore, at $t = t_{j+1}$, we can obtain from (10) that

$$\begin{aligned} V(t_{j+1}) &\leq [\alpha + \beta e^{\lambda(\bar{\tau}+r)} + \omega r e^{\lambda r}] Me^{-\lambda(t_{j+1}-t_0)} \\ &= e^{-(\lambda+c)(t_{j+2}-t_{j+1})} Me^{-(\lambda+c)(t_{j+1}-t_0)} e^{c(t_{j+1}-t_0)} \\ &= M e^{-(\lambda+c)(t_{j+2}-t_0)} e^{c(t_{j+1}-t_0)}. \end{aligned}$$

When $t \in (t_{j+1}, t_{j+2})$, from (7),

$$\begin{aligned} V(t) &\leq M e^{-(\lambda+c)(t_{j+2}-t_0)} e^{c(t_{j+1}-t_0)} e^{c(t-t_{j+1})} \\ &= M e^{-(\lambda+c)(t_{j+2}-t_0)} e^{c(t-t_0)}. \end{aligned}$$

This proves that (12) is true for $k = j + 1$, which implies that (12) holds by the mathematical induction method. Therefore,

$$\begin{aligned} V_1(t) &\leq V(t) \leq M e^{-(\lambda+c)(t_{k+1}-t_0)} e^{c(t_{k+1}-t_0)} \\ &= M e^{-\lambda(t_{k+1}-t_0)}, \quad t \in [t_k, t_{k+1}), \quad k \geq p. \end{aligned}$$

When $k \rightarrow \infty, t \rightarrow \infty, M e^{-\lambda(t_{k+1}-t_0)} \rightarrow 0, V_1(t) \rightarrow 0$, which implies that for any $i = 1, 2, \dots, N, |x_i(t) - Ave(x(0))| \rightarrow 0$ as $t \rightarrow \infty$. On the other hand, for any $i, j = 1, 2, \dots, N$, we have that

$$|x_i(t) - x_j(t)| \leq |x_i(t) - Ave(x(0))| + |x_j(t) - Ave(x(0))| \leq 2 \max_{1 \leq i \leq N} |x_i(t) - Ave(x(0))|.$$

By taking limits $t \rightarrow \infty$ on both sides of the above inequality, we can conclude that $\lim_{t \rightarrow \infty} |x_i(t) - x_j(t)| = 0$ for any $i, j = 1, 2, \dots, N$, hence protocol (2) leads to the consensus for agents in complex-valued multi-agent system (1). \square

Remark 1 The parameter $\omega \in (0, 1]$ can adjust the value of $\alpha + \beta + \omega r$ to guarantee that $\alpha + \beta + \omega r < 1$ such that condition (6) in Theorem 1 can be satisfied for relatively large delay size of r . It can be seen from the proof of Theorem 1 that the impulsive part of protocol (2) plays control effect to accelerate the consensus process, while the continuous-time part of protocol (2) may either accelerate or decelerate such process, and condition (6) implies that impulsive distances have to be suitably designed such that hybrid protocol (2) can solve the consensus problem.

Theorem 2 Suppose \mathcal{G}_A is balanced, \mathcal{G}'_A is balanced and strongly connected. Let $\lambda_2(\mathcal{L}'_s)$ be the second smallest eigenvalue of $\mathcal{L}'_s = \frac{\mathcal{L}' + \mathcal{L}'^T}{2}$, and denotes $\rho_{\min} = (\sqrt{1 - 2\lambda_2(\mathcal{L}'_s)} + \|\mathcal{L}'\|^2 + \bar{\tau} \|\mathcal{L}'\| \|\mathcal{L}'\|)^2$. If $\rho_{\min} < 1$ and constant impulsive distance T satisfying

$$\bar{\tau} < T < \begin{cases} -\frac{\ln(\rho_{\min} + r)}{\|\mathcal{L}'\|^2}, & \text{if } 0 < r < u^* - \rho_{\min}, \\ \frac{1 - \rho_{\min}}{\|\mathcal{L}'\|^2 \rho_{\min}}, & \text{if } u^* - \rho_{\min} \leq r < \infty, \end{cases} \quad (14)$$

where $u^* = e^{W(\rho_{\min} e)} - 1$ and $W(\cdot)$ is the Lambert W function, then protocol (2) leads to the consensus for agents in complex-valued multi-agent system (1).

Proof It can be seen from Theorem 1 that α, β depend on parameter ε , and c depends on parameter ω , where $0 < \omega \leq 1$. If we define $\rho = \rho(\varepsilon) =$

$\alpha + \beta = (1 + \varepsilon)(1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2) + (1 + \frac{1}{\varepsilon})(\bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|)^2$, condition (6) in Theorem 1 implies that the consensus result will be achieved if $T < \frac{-\omega \ln(\rho + \omega r)}{\|\mathcal{L}'\|^2}$. In order to find the upper bound for constant impulsive distances T , we will specify the values of parameters ε and ω to maximize $-\omega \ln(\rho + \omega r)$. For any given $0 < \omega \leq 1$, define $F(\rho) = -\omega \ln(\rho + \omega r)$. In order to maximize $F(\rho)$, we need $\rho + \omega r < 1$, hence $\rho < 1 - \omega r$. By applying the extreme value theory, $\rho = \rho(\varepsilon)$ attains its minimum when $\varepsilon = \frac{\bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|}{\sqrt{1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2}}$, and $\rho_{\min} = \min_{\varepsilon > 0} \rho = (\sqrt{1 - 2\lambda_2(\mathcal{L}'_s) + \|\mathcal{L}'\|^2} + \bar{\tau}\|\mathcal{L}\|\|\mathcal{L}'\|)^2$. If $\rho_{\min} < 1 - \omega r < 1$, then

$$F(\rho) = -\omega \ln(\rho + \omega r), \quad \rho \in [\rho_{\min}, 1 - \omega r),$$

and for $\rho \in [\rho_{\min}, 1 - \omega r)$, $F'(\rho) = \frac{-\omega}{\rho + \omega r} < 0$. Therefore, $\max F(\rho) = F(\rho_{\min}) = -\omega \ln(\rho_{\min} + \omega r)$. Next, we define $G(\omega) = \frac{-\omega \ln(\rho_{\min} + \omega r)}{\|\mathcal{L}'\|^2}$, where $0 < \omega < \frac{1 - \rho_{\min}}{r}$ and $0 < \omega \leq 1$. We have to consider the following two cases depending on the size of delay r .

Case 1: if $0 < \omega < \frac{1 - \rho_{\min}}{r} \leq 1$, then $1 - \rho_{\min} \leq r < \infty$, and $G(\omega) = \frac{-\omega \ln(\rho_{\min} + \omega r)}{\|\mathcal{L}'\|^2}$, $\omega \in (0, \frac{1 - \rho_{\min}}{r})$. Define $u = \rho_{\min} + \omega r$, where $\omega \in (0, \frac{1 - \rho_{\min}}{r})$. Then, $G'(\omega) = 0$ implies $\ln u - \frac{\rho_{\min}}{u} + 1 = 0$, $u \in (\rho_{\min}, 1)$. Next, we define

$$f(u) = \ln u - \frac{\rho_{\min}}{u} + 1, \quad u \in (\rho_{\min}, 1), \quad (15)$$

then, $f(\rho_{\min}) = \ln(\rho_{\min}) < 0$, $f(1) = 1 - \rho_{\min} > 0$, and $f'(u) = \frac{1}{u} + \frac{\rho_{\min}}{u^2} > 0$ for $u \in (\rho_{\min}, 1)$. Then by IVT, there exists a unique $u^* \in (\rho_{\min}, 1)$ such that $f(u^*) = 0$. Let $v = \ln u$, and $u = e^v$, then $f(u) = 0$ implies $e^{v+1}(v+1) = \rho_{\min} \cdot e$. Therefore, based on the property of the Lambert W function, we have $v = W(\rho_{\min}e) - 1$, and $u^* = e^v = e^{W(\rho_{\min}e) - 1} \in (\rho_{\min}, 1)$. Therefore, there exists a unique $\omega^* = \frac{u^* - \rho_{\min}}{r}$ such that $G'(\omega^*) = 0$, where $0 < \omega^* = \frac{u^* - \rho_{\min}}{r} < \frac{1 - \rho_{\min}}{r} \leq 1$. According to (15), if $0 < \omega < \omega^*$, $\rho_{\min} < u < u^*$, then $f(u) < 0$ and $G'(\omega) > 0$, and if $\omega^* < \omega < \frac{1 - \rho_{\min}}{r}$, $u^* < u < 1$, then $f(u) > 0$ and $G'(\omega) < 0$. Therefore, when $1 - \rho_{\min} \leq r < \infty$,

$$\max_{\omega \in (0, \frac{1 - \rho_{\min}}{r})} G(\omega) = G(\omega^*) = \frac{(u^* - \rho_{\min})^2}{\|\mathcal{L}'\|^2 r u^*} \leq \frac{1 - \rho_{\min}}{\|\mathcal{L}'\|^2 \rho_{\min}}.$$

Case 2: if $0 < \omega \leq 1 < \frac{1 - \rho_{\min}}{r}$, then $0 < r < 1 - \rho_{\min}$, and $G(\omega) = \frac{-\omega \ln(\rho_{\min} + \omega r)}{\|\mathcal{L}'\|^2}$, $\omega \in (0, 1]$. Let $u = \rho_{\min} + \omega r$, since $\omega \in (0, 1]$, then $u \in (\rho_{\min}, \rho_{\min} + r]$. Define

$$h(u) = \ln u - \frac{\rho_{\min}}{u} + 1, \quad u \in (\rho_{\min}, \rho_{\min} + r]. \quad (16)$$

Then, $G'(\omega) = 0$ implies $h(u) = 0$. Moreover, we can obtain $h(\rho_{\min}) = \ln(\rho_{\min}) < 0$, and $h'(u) = \frac{1}{u} + \frac{\rho_{\min}}{u^2} > 0$ for $u \in (\rho_{\min}, \rho_{\min} + r]$. From previous discussion,

we know that $h(u^*) = 0$ when $u^* = e^{W(\rho_{\min}e)-1}$. Therefore, if $u^* \in (\rho_{\min}, \rho_{\min} + r]$, which implies $u^* - \rho_{\min} \leq r < 1 - \rho_{\min}$, then there exists a unique $\omega^* = \frac{u^* - \rho_{\min}}{r} \in (0, 1]$ such that $G'(\omega^*) = 0$. According to (16), if $0 < \omega < \omega^*$, $\rho_{\min} < u < u^*$, then $h(u) < 0$ and $G'(\omega) > 0$, and if $\omega^* < \omega \leq 1$, $u^* < u \leq \rho_{\min} + r$, then $h(u) > 0$ and $G'(\omega) < 0$. Therefore, if $u^* - \rho_{\min} \leq r < 1 - \rho_{\min}$, then

$$\max_{\omega \in (0,1]} G(\omega) = G(\omega^*) = \frac{(u^* - \rho_{\min})^2}{\|\mathcal{L}\|^2 r u^*} \leq \frac{1 - \rho_{\min}}{\|\mathcal{L}\|^2 \rho_{\min}}.$$

If $u^* > \rho_{\min} + r$, then $0 < r < u^* - \rho_{\min}$, and $\omega^* = \frac{u^* - \rho_{\min}}{r} > 1$, according to (16), when $0 < \omega \leq 1$, $\rho_{\min} < u \leq \rho_{\min} + r$, then $h(u) < 0$ and $G'(\omega) > 0$ for all $\omega \in (0, 1]$. Therefore, when $0 < r < u^* - \rho_{\min}$,

$$\max_{\omega \in (0,1]} G(\omega) = G(1) = \frac{-\ln(\rho_{\min} + r)}{\|\mathcal{L}\|^2}.$$

According to the above discussion, we can conclude that protocol (2) leads to consensus for agents in complex-valued multi-agent system (1) if impulsive distances satisfy (14), which completes the proof. \square

Remark 2 Comparing with the consensus results established in [9], Theorem 2 provides a delay- r dependent condition to find the upper bound for the length of impulsive intervals based on various delay sizes of r such that protocol (2) solves the complex-valued consensus problem, the consensus result will always be achieved as long as the designed impulsive distances are less than the upper bound obtained in (14).

4 Numerical Simulations

Consider complex-valued multi-agent system (1) with hybrid protocol (2) consisting of eight agents. In Fig. 1, the solid lines denote the edges of fixed digraphs \mathcal{G}_A at continuous-time intervals, and the dashed lines represent the edges of fixed digraphs \mathcal{G}'_A at each impulsive instant. In Example 1, at each continuous-time interval, \mathcal{G}_A is assumed to have weights 0.12 between the 2nd agent and the 8th agent; weights 0.15 between the 3rd agent and the 7th agent, and weights 0.18 between the 4th agent and the 6th agent; digraph \mathcal{G}'_A has equal weight 0.08 at each impulsive instant.

Example 1 Consider hybrid impulsive protocol (2) with fixed digraphs \mathcal{G}_A and \mathcal{G}'_A shown in Fig. 1. It can be seen that \mathcal{G}_A is balanced with $\|\mathcal{L}\| = 0.36$, \mathcal{G}'_A is balanced and strongly connected with $\|\mathcal{L}'\| = 0.16$, and $\lambda_2(\mathcal{L}'_s) = 0.0234$. If $\bar{\tau} = 0.03$, then $\rho_{\min} = 0.982 < 1$, and $u^* = 0.991$. For $r = 0.006$, $r < u^* - \rho_{\min}$. Theorem 2 implies that protocol (2) can solve the consensus problem if the uniform impulsive distance T satisfies $\bar{\tau} < T < -\frac{\ln(\rho_{\min} + r)}{\|\mathcal{L}'\|^2} = 0.093$. The initial states are chosen as

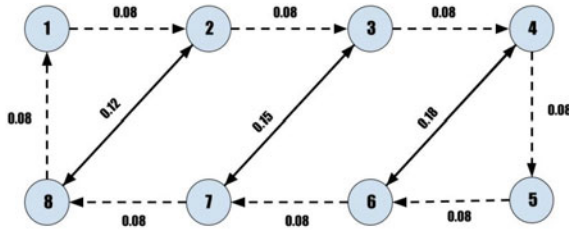


Fig. 1 Fixed topologies \mathcal{G}_A and \mathcal{G}'_A

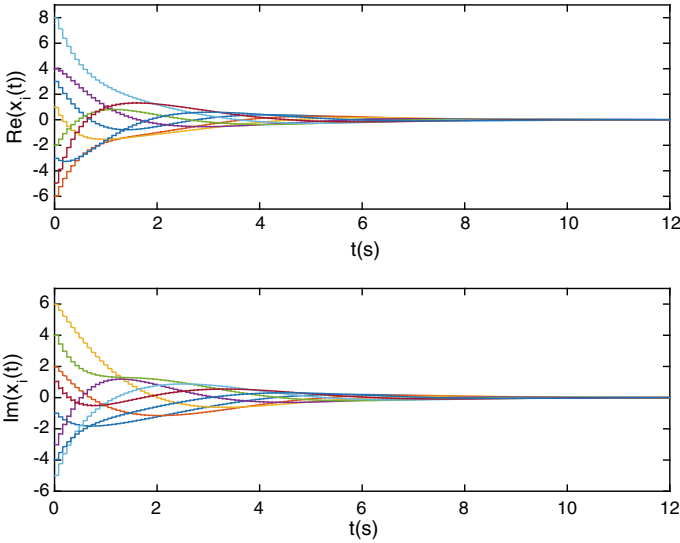


Fig. 2 Average consensus process for $\bar{\tau} = 0.03, r = 0.006$, and $T = 0.08$

$x(0) = [3 - i, -6 + 2i, 1 + 6i, 4 - 3i, -2 + 4i, 8 - 5i, -5 + i, -3 - 4i]^T$, and if we design $T = 0.08 < 0.093$, then all the agent states reach a consent. The consensus process for both real and imaginary parts of the state of agents are shown in Fig. 2, and the final consent state is nothing but $Ave(x(0)) = 0$. For $r = 2, r \geq u^* - \rho_{\min}$, then consensus can be achieved if $\bar{\tau} < T < \frac{1 - \rho_{\min}}{\|\mathcal{L}\|^2 \rho_{\min}} = 0.141$ based on Theorem 2. With the same initial conditions, choose $T = 0.12 < 0.141$, Fig. 3 shows that the consensus result can still be confirmed even if $r > T$.

5 Conclusion

We have studied the consensus problem of complex-valued multi-agent systems. A hybrid impulsive consensus protocol that takes into account both the fixed network

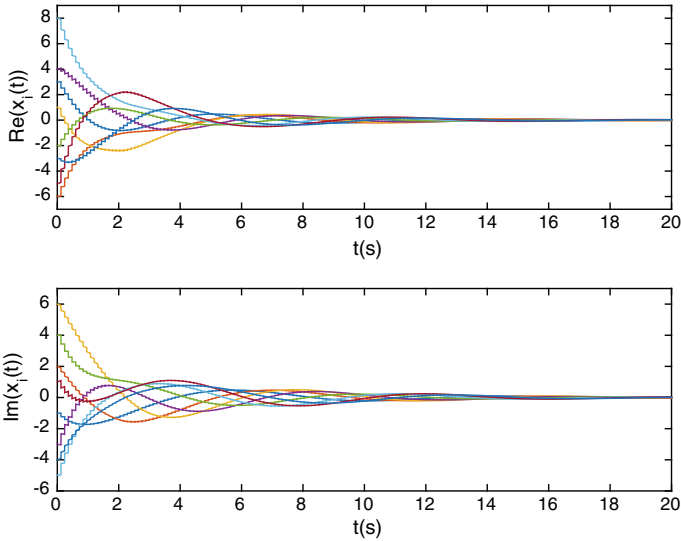


Fig. 3 Average consensus process for $\bar{\tau} = 0.03$, $r = 2$, and $T = 0.12$

topologies and time delays has been proposed. Delay dependent sufficient conditions have been derived to guarantee that the consensus result can be achieved in complex space via the proposed consensus protocol by employing the method of Lyapunov functionals. Our result has shown that the proposed consensus protocols can lead complex-valued multi-agent systems to achieve consensus if one can construct appropriate network topologies and suitably design the length of impulsive intervals based on various delay sizes for the continuous-time part of the consensus protocol. Numerical simulations have been provided to demonstrate the effectiveness of the obtained theoretical results.

References

1. Bliman, P., Ferrari-Trecate, G.: Average consensus problems in networks of agents with delayed communications. *Automatica* **44**, 1985–1995 (2008)
2. Satunin, S., Babkin, E.: A multi-agent approach to intelligent transportation systems modeling with combinatorial auctions. *Expert Syst. with Appl.* **41**, 6622–6633 (2014)
3. Shen, K., Jing, Z., Dong, P.: Simultaneous target tracking and sensor location refinement in distributed sensor networks. *Signal Process.* **153**, 123–131 (2018)
4. Olfati-Saber, R., Murray, R.: Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans. Autom. Control* **49**(9), 1520–1533 (2004)
5. Lin, P., Jia, Y.: Average consensus in networks of multi-agents with both switching topology and coupling time-delay. *Physica A* **387**, 303–313 (2008)
6. Sun, Y., Wang, L.: Consensus of multi-agent systems in directed networks with nonuniform time-varying delays. *IEEE* **54**(7), 1607–1613 (2009)

7. Jiang, H., Yu, J., Zhou, C.: Consensus of multi-agent linear dynamic systems via impulsive control protocols. *Int. J. Syst. Sci.* **42**, 967–976 (2011)
8. Liu, X., Zhang, K., Xie, W.: Consensus seeking in multi-agent systems via hybrid protocols with impulse delays. *Nonlinear Anal.: Hybrid Syst.* **25**, 90–98 (2017)
9. Liu, X., Zhang, K., Xie, W.: Consensus of multi-agent systems via hybrid impulsive protocols with time-delay. *Nonlinear Anal.: Hybrid Syst.* **30**, 134–146 (2018)
10. Gibbon, J., McGuinness, M.: The real and complex Lorenz equations in rotating fluids and lasers. *Physica D* **5**(1), 108–122 (1982)
11. Bolognani, S., et al.: Adaptive output feedback control for control for complex-valued reaction-advection-diffusion systems. In: *Proceedings of the 2008 American Control Conference*, pp. 961–966 (2008)
12. Xu, X., Zhang, J., Shi, J.: Dynamical behaviour analysis of delayed complex-valued neural networks with impulsive effect. *Int. J. Syst. Sci.* **48**, 686–694 (2017)
13. Xu, W., Zhu, S., et al.: Adaptive synchronization of memristor-based complex-valued neural networks with time delays. *Neurocomputing* **364**, 119–128 (2019)
14. Zhang, H., Wang, X., et al.: Combination synchronization and stability analysis of time-varying complex-valued neural networks. *Chaos, Solitons and Fractals* **131**, 109485 (2020)

Input-to-State Stability for Delayed Hybrid Systems and H_∞ Control



Taghreed G. Sugati, Mohamad S. Alwan, and Xinzhi Liu

Abstract This paper addresses the problems of input-to-state stability/stabilization (ISS) and designing a robust reliable H_∞ control for a class of switched systems with state delay and time-varying, bounded disturbing input. The methodology of Razumikhin with multiple Lyapunov functions is used to establish the ISS property. The importance of this method is that it provides delay-independent sufficient conditions to guarantee the ISS of the system modes, and later this result will be applied to design the feedback H_∞ controller not only when all the actuators are operational, but also when some of them experience failure. The non-zero output of faulty actuators are treated as a disturbance signal that is augmented with the system disturbance input. The organization of mode switching is ruled by the general framework of average dwell-time (ADT) switching law. Finally, the proposed theoretical results are clarified by a numerical example with simulations.

Keywords Hybrid systems · Lyapunov-Razumikhin method · Switched dynamics · Feedback controller design · Average-dwell-time

1 Introduction

A switched system is a special class of hybrid systems that consists of a family of continuous- or discrete-time dynamical subsystems, and a switching signal that organizes the switchings among the system modes. The importance of studying

T. G. Sugati (✉)
University of King Abdulaziz, Jeddah, Saudi Arabia
e-mail: tsogati@kau.edu.sa

M. S. Alwan
University of Saskatchewan, Saskatoon S7N 5E6, Canada
e-mail: m.alwan@math.usask.ca

X. Liu
University of Waterloo, Waterloo N2L 3G1, Canada
e-mail: xzliu@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_21

221

switched systems is many folds. It is a proper tool to cover many natural and human-made phenomena [1–4], many dynamical systems can only be stabilized by several control laws, but not by one law [7], it can be used to reduce the complexity of many sophisticated systems [5], and it is well known that the stability of such a system is not guaranteed by the stability of the individual modes unless the switching among them is ruled by a logic-based switching signal. As a result, switched systems have received a large amount of literature including books [6–8] and special issues [9–12].

A reliable control is meant to be a controller that tolerates failures in its components, particularly actuators or sensors. In reality, such failures are frequently encountered and an immediate repair may be impossible in some critical cases as in the case of aerospace vehicles or submarine systems, etc [14–21].

The ISS notion, introduced in [22], has been realized to be an efficient tool to deal with such disturbances. Briefly, the ISS addresses the system response to a bounded disturbance when the unforced system is asymptotically stable [22–24].

The novelty here is to address the ISS for the switched time-delayed system and design robust H_∞ reliable controller with nonlinear perturbation by using the Lyapunov-Razumikhin methodology. The jump among the system modes is controlled by the ADT switching signal. Later, these results are applied to switched control systems with possible faulty actuators. The actuator output signal is treated as a disturbance signal augmented with the system input disturbance.

2 Problem Formulation

Consider the nonlinear switched system

$$\begin{cases} \dot{x} = f_{\varrho(t)}(x_t, w(t)), \\ x_{t_0}(s) = \phi(s), \quad s \in [-r, 0], \quad r > 0, \end{cases} \quad (1)$$

where $x \in \mathbb{R}^n$ is the system state and $w \in \mathbb{R}^p$ is an input disturbance, which is assumed to be in $L_2[t_0, \infty)$, that is $\|w\|_2^2 = \int_{t_0}^\infty \|w(t)\|^2 dt < \infty$. For all $t \in \mathbb{R}_+ = [0, \infty)$, let $x(t)$ be a function defined on $[t_0, \infty)$. Then, we define the function $x_t : [-r, 0] \rightarrow \mathbb{R}^n$ by $x_t(s) = x(t + s)$ for all $s \in [-r, 0]$, and its norm by $\|x_t\|_r = \sup_{t-r \leq \theta \leq t} \|x(\theta)\|$, where $r > 0$ is the time delay. ϱ is the switching rule which is a piecewise constant function defined by $\varrho : [t_0, \infty) \rightarrow \mathcal{S} = \{1, 2, \dots, N\}$, for a natural number N .

Definition 1 [23] A function $\alpha \in \mathcal{C}([0, a], \mathbb{R}_+)$ is said to be in class \mathcal{K} if $\alpha(0) = 0$ and it is strictly increasing.

Definition 2 [23] System (1) is said to be globally exponentially ISS if there exist $\lambda > 0, \bar{\lambda} > 0$ and a function $\rho \in \mathcal{K}$ such that $x(t)$ exists $\forall t \geq t_0$ and satisfies

$$\|x\| \leq \bar{\lambda} \|x_{t_0}\|_r e^{-\lambda(t-t_0)} + \rho \left(\sup_{t_0 \leq \tau \leq t} \|w(\tau)\| \right).$$

Lemma 1 For any $\xi_j > 0$ ($j = 1, 2, 3$) and a positive-definite matrix P , we have

- (i) $2x^T P G w \leq x^T (\xi_1 P G G^T P) x + \frac{1}{\xi_1} w^T w$. Moreover, for $x \in \mathcal{C}_r$, if $\|x(t-r)\|_r^2 \leq q \|x\|^2$ with $q > 1$, then
- (ii) $2x^T P \bar{A} x(t-r) \leq x^T (\xi_2 P \bar{A} (\bar{A})^T P + \frac{q}{\xi_2} I) x$.
- (iii) $2x^T P f(x(t)) \leq x^T (\xi_3 P^2 + \frac{1}{\xi_3} \delta I) x$, where $\delta > 0$ such that $\|f(x(t))\|_r^2 \leq \delta \|x(t)\|^2$.

Average Dwell Time [13]. The number of switches $N(t_0, t)$ in (t_0, t) for a finite t satisfies $N(t_0, t) \leq N_0 + \frac{t-t_0}{\tau_a}$, where N_0 is the chatter bound, and τ_a is the ADT.

3 Main Results

The following theorem gives sufficient conditions of global exponential ISS property of the system, where we use the Lyapunov-Razumikhin method [25].

Theorem 1 For $i \in \mathcal{S}$, let γ be a differentiable \mathcal{X} function. Assume there are positive constants c_1, c_2, r, β , and a function $V_i \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}_+)$ such that

- (i) $c_1 \|x\|^2 \leq V_i(x) \leq c_2 \|x\|^2$ for all $t \geq t_0 - r$;
- (ii) $\dot{V}_i(\psi(0)) < -\lambda V_i(\psi(0))$ whenever the relations $V_i(\psi(s)) \leq q V_i(\psi(0))$ and $\gamma\left(\sup_{t_0 \leq \theta \leq t_k} |w(\theta)|\right) \leq V_i(\psi(0))$ hold for $\psi \in \mathcal{C}_r, s \in [-r, 0]$ and $t \in [t_{k-1}, t_k)$, where $q = \max\{\mu h, e^{\lambda r}\} > 1$ with $\mu = c_2/c_1 \geq 1$;
- (iii) for all $k, r \leq t_k - t_{k-1} \leq \beta$ and the ADT condition holds;
- (iv) for $s \in [-r, 0]$ and $h > 1, V_i(x(t+s)) \leq h V_j(x(t))$ for any $i, j \in \mathcal{S}$ and all $t \geq t_0$.

Then, system (1) is globally exponentially ISS.

Proof Let $x(t) = x(t, t_0, \phi)$ be any solution of system (1) with $x_{t_0} = \phi$ and $v_i(t) = V_i(x(t))$. First, we want to show that every mode is globally exponentially ISS using conditions (i) and (ii). For any $i \in \mathcal{S}$, and $k \in \mathbb{N}, t \in [t_{k-1}, t_k)$, we shall show that

$$v_i(t) \leq c_2 \|x_{t_{k-1}}\|_r^2 e^{-\lambda(t-t_{k-1})} + \gamma\left(\sup_{t_0 \leq s \leq t} \|w(s)\|\right). \quad (2)$$

Define

$$Q_i(t) = \begin{cases} v_i(t) - c_2 \|x_{t_{k-1}}\|_r^2 e^{-\lambda(t-t_{k-1})} - \gamma\left(\sup_{t_0 \leq s \leq t} \|w(s)\|\right), & t \in [t_{k-1}, t_k), k \in \mathbb{N} \\ v_i(t) - c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t-t_0)}, & t \in [t_0 - r, t_0). \end{cases}$$

We will show that $Q_i(t) \leq 0$ for all $t \geq t_0 - r$. For $t \in [t_0 - r, t_0]$, it is clear that $Q_i(t) \leq 0$. By condition (i),

$$v_i(t) \leq c_2 \|x\|^2 \leq c_2 \|x_{t_0}\|_r^2 \leq c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t-t_0)} \tag{3}$$

as $-\lambda(t - t_0) > 0$ for $t \in [t_0 - r, t_0]$. So, we have $Q_i(t) = v_i(t) - c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t-t_0)} \leq 0$. Step 1, for $t \in [t_0, t_1]$, we need to show

$$Q_i(t) = v_i(t) - c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t-t_0)} - \gamma \left(\sup_{t_0 \leq \theta \leq t_1} \|w(\theta)\| \right) \leq 0. \tag{4}$$

For any $i \in \mathcal{S}$, let $\alpha_i > 0$ be arbitrary, and we show $Q_i(t) \leq \alpha_i$ for $[t_0, t_1]$. If not, then there would exist some $t \in [t_0, t_1]$ so that $Q_i(t) > \alpha_i$. Let

$$t_i^* = \inf\{t \in [t_0, t_1] : Q_i(t) > \alpha_i, i \in \mathcal{S}\}.$$

We also have $Q_i(t_0) \leq v_i(t_0) - c_2 \|x_{t_0}\|_r^2 \leq c_2 (\|x(t_0)\|^2 - \|x_{t_0}\|_r^2) \leq 0$. Since $Q_i(t) \leq 0 < \alpha_i$ for $t \in [t_0 - r, t_0]$, then $t_i^* \in (t_0, t_1)$. Also, since $Q_i(t)$ is continuous on $[t_0, t_1]$, we have $Q_i(t_i^*) = \alpha_i$ and $Q_i(t) \leq \alpha_i$ for $[t_0 - r, t_i^*]$. So, we have

$$v_i(t_i^*) = Q_i(t_i^*) + c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*-t_0)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \tag{5}$$

and for $s \in [-r, 0]$, we have

$$\begin{aligned} v_i(t_i^* + s) &= Q_i(t_i^* + s) + c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*+s-t_0)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*+s} \|w(\theta)\| \right) \\ &\leq \alpha_i + c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*-t_0)} e^{\lambda r} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \\ &\leq [\alpha_i + c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*-t_0)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right)] e^{\lambda r} \\ &= e^{\lambda r} v_i(t_i^*) \leq q v_i(t_i^*), \end{aligned} \tag{6}$$

where from (5), we have used $\gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \leq v_i(t_i^*)$. Thus, from condition (ii), we have $\dot{v}_i(t_i^*) \leq -\lambda v_i(t_i^*)$ which implies that

$$\begin{aligned} \dot{Q}_i(t_i^*) &= \dot{v}_i(t_i^*) + \lambda c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*-t_0)} - \dot{\gamma} \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \\ &\leq -\lambda \left[v_i(t_i^*) - c_2 \|x_{t_0}\|_r^2 e^{-\lambda(t_i^*-t_0)} - \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \right] = -\lambda \alpha_i. \end{aligned} \tag{7}$$

Therefore, $Q_i(t)$ is decreasing at t_i^* which is a contradiction for being increasing at t^* according to the definition of t^* . Thus, we get $Q_i(t) \leq \alpha_i$ for all $t \in [t_0, t_1]$. Let $\alpha_i \rightarrow 0^+$, then we have $Q_i(t) \leq 0$ for $t \in [t_0, t_1]$.

Step 2, for any $i \in \mathcal{S}$ assume $Q_i(t) \leq 0$ for all $t \in [t_{k-1}, t_k)$ for $k = 1, \dots, m$.

$$\begin{aligned} Q_i(t_m) &= v_i(t_m) - c_2 \|x_{t_m}\|_r^2 - \gamma \left(\sup_{t_0 \leq \theta \leq t_{m+1}} \|w(\theta)\| \right) \\ &\leq c_2 (\|x(t_m)\|^2 - \|x_{t_m}\|_r^2) - \gamma \left(\sup_{t_0 \leq \theta \leq t_{m+1}} \|w(\theta)\| \right) \leq 0. \end{aligned}$$

Step 3, we will show that $Q_i(t) \leq 0$ for all $t \in [t_m, t_{m+1})$, i.e., we aim to show that

$$v_i(t) \leq c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t-t_m)} + \gamma \left(\sup_{t_0 \leq s \leq t} \|w(s)\| \right).$$

To do so, we need first to prove that $Q_i(t) \leq \alpha_i$ for $t \in [t_m, t_{m+1})$ and $i \in \mathcal{S}$. If this were not true, then there would be $t \in [t_m, t_{m+1})$ such that for $i \in \mathcal{S}$, $Q_i(t) > \alpha_i$. Let

$$t_i^* = \inf\{t \in [t_m, t_{m+1}) : Q_i(t) > \alpha_i, i \in \mathcal{S}\}.$$

Then, by the continuity, we have $Q_i(t_i^*) = \alpha_i$ and $Q_i(t) \leq \alpha_i$ for all $t \in [t_m, t_i^*)$, i.e., $\dot{Q}_i(t_i^*) > 0$. Thus, we have

$$v_i(t_i^*) = \alpha_i + c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t_i^*-t_m)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right). \quad (8)$$

We want to show $v_i(t_i^* + s) \leq v_i(t_i^*)$ for $s \in [-r, 0]$.

Case 1. If $t_i^* + s \in [t_m, t_{m+1})$, then we have for each $i \in \mathcal{S}$

$$\begin{aligned} v_i(t_i^* + s) &= Q_i(t_i^* + s) + c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t_i^*+s-t_m)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*+s} \|w(\theta)\| \right) \\ &\leq \left[\alpha_i + c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t_i^*-t_m)} + \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \right] e^{\lambda s} \\ &= e^{\lambda s} v_i(t_i^*) \leq q v_i(t_i^*). \end{aligned} \quad (9)$$

Case 2. If $t_i^* + s \in [t_m - r, t_m)$. Then, for $i, j \in \mathcal{S}$ and $t \geq t_0 - r$, we have $v_i(t) \leq \mu v_j(t)$ with $\mu = c_2/c_1 \geq 1$. Using (iv) implies that $v_i(t_i^* + s) \leq \mu v_j(t_i^* + s) \leq \mu h v_i(t_i^*) \leq q v_i(t_i^*)$, where $q = \max\{e^{\lambda r}, \mu h\}$. Also, from (8), we have $\gamma(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\|) \leq v_i(t_i^*)$. Thus, from condition (ii), we have $\dot{v}_i(t_i^*) \leq -\lambda v_i(t_i^*)$ which implies

$$\begin{aligned} \dot{Q}_i(t_i^*) &= \dot{v}_i(t_i^*) + \lambda c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t_i^*-t_m)} - \dot{\gamma} \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \\ &\leq -\lambda \left[v_i(t_i^*) - c_2 \|x_{t_m}\|_r^2 e^{-\lambda(t_i^*-t_m)} - \gamma \left(\sup_{t_0 \leq \theta \leq t_i^*} \|w(\theta)\| \right) \right] = -\lambda \alpha_i. \end{aligned} \quad (10)$$

So that, $Q_i(t)$ is decreasing at t_i^* which is a contradiction for being increasing at t_i^* according to the definition of t_i^* . Thus, we get $Q_i(t) \leq \alpha_i$ for all $t \in [t_m, t_{m+1})$. Let $\alpha_i \rightarrow 0^+$, then we have $Q_i(t) \leq 0$ for $t \in [t_m, t_{m+1})$. By induction, we conclude that $Q_i(t) \leq 0$ for all $t \geq t_0 - r$. Thus, we have proved that, for $t \in [t_{k-1}, t_k)$,

$$v_i(t) \leq c_2 \|x_{t_{k-1}}\|_r^2 e^{-\lambda(t-t_{k-1})} + \gamma \left(\sup_{t_0 \leq s \leq t} \|w(s)\| \right). \tag{11}$$

By (i), we get $\|x\| \leq \sqrt{\mu} \|x_{t_{k-1}}\|_r e^{-\lambda(t-t_{k-1})/2} + \sqrt{\frac{1}{c_1} \gamma \left(\sup_{t_0 \leq s \leq t} \|w(s)\| \right)}$. This proves that every mode is ISS. Second, we aim to show the switched system is ISS. Since condition (i) is assumed to hold for all $t \geq t_0 - r$, then we have from (11)

$$V_i(x(t)) \leq \mu V_i(x(t_{k-1} - r)) e^{-\lambda(t-t_{k-1})} + \gamma \left(\sup_{t_0 \leq s \leq t} \|w(s)\| \right). \tag{12}$$

For any $i \in \mathcal{S}$ and all $t \in [t_{k-1}, t_k)$, we have

$$\begin{aligned} V_i(x(t)) &\leq \mu^{2k-1} e^{(k-1)\lambda r} e^{-\lambda(t-t_0)} V_i(x_{t_0}) + \left(\sum_{j=0}^{k-1} (\mu^2)^j \right) \gamma \left(\sup_{t_0 \leq s \leq t_k} \|w(s)\| \right) \\ &\leq \mu^k (\mu e^{\lambda r})^{k-1} e^{-\lambda(t-t_0)} V_i(x_{t_0}) + k(\mu^2)^{k-1} \gamma \left(\sup_{t_0 \leq s \leq t_k} \|w(s)\| \right) \\ &\leq (\mu \varrho)^k \varrho^{-1} e^{-\lambda(t-t_0)} V_i(x_{t_0}) + k(\mu^2)^{k-1} \gamma \left(\sup_{t_0 \leq s \leq t_k} \|w(s)\| \right) \\ &\leq e^{k \ln(\mu \varrho) - \ln(\varrho) - \lambda(t-t_0)} V_i(x_{t_0}) + \Gamma(t), \end{aligned}$$

where $\varrho = \mu e^{\lambda r}$ and $\Gamma(t) = k(\mu^2)^{k-1} \gamma \left(\sup_{t_0 \leq s \leq t_k} \|w(s)\| \right)$. Using the ADT with $N_0 = \frac{\eta}{\ln(\mu \varrho)}$, $\tau_a = \frac{\ln(\mu \varrho)}{\lambda - \nu}$, ($0 < \nu < \lambda$), for an arbitrary positive constant η , we get

$$V_i(x(t)) \leq e^{\eta + \ln \mu - \nu(t-t_0)} V_i(x_{t_0}) + \Gamma(t) \leq D e^{-\nu(t-t_0)} \|x_{t_0}\|_r^2 + \Gamma(t),$$

where $D = c_2 \mu e^\eta$. This implies that $\|x\| \leq b \|x_{t_0}\|_r e^{-\nu(t-t_0)/2} + \bar{\gamma}(t)$, $t \geq t_0$, where $b = \mu \sqrt{e^\eta}$, and $\bar{\gamma}(t) = \sqrt{\Gamma(t)/c_1}$ is class \mathcal{L} . The proof is complete.

3.1 Linear Systems

Consider the switched input/output linear system with time delay

$$\begin{cases} \dot{x} = A_{\varrho(t)} x + \bar{A}_{\varrho(t)} x(t-r) + B_{\varrho(t)} u + G_{\varrho(t)} w + f_{\varrho(t)}(x(t)), \\ z = C_{\varrho(t)} x + F_{\varrho(t)} u, \\ x_{t_0}(s) = \phi(s), \quad s \in [-r, 0], \quad r > 0, \end{cases} \tag{13}$$

where $u \in \mathbb{R}^l$ is the control input and $z \in \mathbb{R}^r$ is the controlled output. A_i is non Hurwitz, $K_i \in \mathbb{R}^{l \times n}$ is the control gain such that $u = K_i x$ and (A_i, B_i) is stabilizable, $f_i(\cdot) \in \mathbb{R}^n$, and A_i, B_i, G_i, C_i, F_i are known matrices. The closed-loop system is

$$\begin{cases} \dot{x} = (A_i + B_i K_i)x + \bar{A}_i x(t-r) + G_i w + f_i(x(t)) \\ z = C_{ic} x, \quad C_{ic} = C_i + F_i K_i. \end{cases} \quad (14)$$

Define $G_{ic} = (G_i \ B_i \sigma)$, then the closed-loop system in the faulty case becomes

$$\dot{x} = (A_i + B_{i\bar{\sigma}} K_i)x + \bar{A}_i x(t-r) + G_{ic} w_\sigma^F + f_i(x(t)). \quad (15)$$

Definition 3 Given a constant $\gamma > 0$, system (13) is said to be ISS- H_∞ if there is a state feedback law $u = K_i x$, such that the closed-loop system (14) is globally exponentially ISS, and the controlled output z satisfies $\|z\|_2^2 = \int_{t_0}^\infty \|z\|^2 dt \leq \gamma^2 \|w\|_2^2 + m_0$, for some $m_0 > 0$.

Corollary 2 For $i \in \mathcal{S}$, let K_i and $\gamma_i > 0$ be given. Assume that there exist positive constants ξ_{ji} ($j = 1, 2, 3$), $\alpha_i > 0$, and a positive-definite matrix P_i satisfying

$$\begin{aligned} & (A_i + B_i K_i)^T P_i + P_i (A_i + B_i K_i) + P_i (\xi_{1i} G_i G_i^T + \xi_{2i} \bar{A}_i (\bar{A}_i)^T + \xi_{3i} I) P_i \\ & + \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I + C_{ic}^T C_{ic} + \alpha_i P_i = 0. \end{aligned} \quad (16)$$

Assume that $\|w\|^2 \leq \xi_{2i} \alpha_i^* V_i(x)$ with $0 < \alpha_i^* < \alpha_i$ and for $k, r \leq t_k - t_{k-1} \leq \beta$ where $\beta > 0$, and the ADT holds. Then, system (14) is globally exponentially ISS- H_∞ .

Proof For $i \in \mathcal{S}$, define $V_i(x) = x^T P_i x$. Then, from condition (ii), we have

$$\begin{aligned} \dot{V}_i(x) &= [(A_i + B_i K_i)x + \bar{A}_i x(t-r) + G_i w + f_i(x(t))]^T P_i x \\ &\quad + x^T P_i [(A_i + B_i K_i)x + \bar{A}_i x(t-r) + G_i w + f_i(x(t))] \\ &\leq x^T [(A_i + B_i K_i)^T P_i + P_i (A_i + B_i K_i) + P_i (\xi_{1i} G_i G_i^T + \xi_{2i} \bar{A}_i (\bar{A}_i)^T \\ &\quad + \xi_{3i} I) P_i + \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I] x + \frac{1}{\xi_{1i}} w^T w \leq -\alpha_i V_i(x) + \frac{1}{\xi_{1i}} w^T w \leq -\lambda_i V_i(x) \\ &\leq -\lambda V_i(x), \end{aligned}$$

where $\lambda_i = \alpha_i - \alpha_i^*$, $\lambda = \min_{i \in \mathcal{S}} \{\lambda_i\}$ and we used Lemma 1, and condition (16). To prove the upper bound to $\|z\|$, let the performance function be defined by $J_i = \int_{t_0}^\infty (z^T z - \gamma_i^2 w^T w) dt$, for $i \in \mathcal{S}$. Then,

$$\begin{aligned}
J_i &= \int_{t_0}^{\infty} (z^T z - \gamma_i^2 w^T w) dt + \int_{t_0}^{\infty} \dot{V}_i dt - V_i(\infty) + V_i(x_0) \\
&\leq \int_{t_0}^{\infty} (z^T z - \gamma_i^2 w^T w) dt + V_i(x_0) + \int_{t_0}^{\infty} \left\{ x^T [(A_i + B_i K_i)^T P_i + P_i (A_i + B_i K_i) \right. \\
&\quad + P_i (\xi_{2i} \bar{A}_i (\bar{A}_i)^T + \xi_{3i} I) P_i + \left. \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I + \gamma_i^{-2} P_i G_i G_i^T P_i - \gamma_i^{-2} P_i G_i G_i^T P_i \right] x \\
&\quad + 2x^T P_i G_i w \left. \right\} dt \\
&= V_i(x_0) + \int_{t_0}^{\infty} \left\{ x^T [(A_i + B_i K_i)^T P_i + P_i (A_i + B_i K_i) + P_i (\xi_{2i} \bar{A}_i (\bar{A}_i)^T \right. \\
&\quad + \xi_{3i} I) P_i + \left. \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I + \gamma_i^{-2} P_i G_i G_i^T P_i + C_{ic}^T C_{ic} \right] x \left. \right\} dt \\
&\quad - \int_{t_0}^{\infty} \gamma_i^2 (w - \gamma_i^{-2} G_i^T P_i x)^T (w - \gamma_i^{-2} G_i^T P_i x) dt.
\end{aligned}$$

Using (16) with $\gamma_i^{-2} = \xi_{1i}$ leads to $J_i \leq V_i(x_0)$; hence, $\|z\|_2^2 \leq \gamma^2 \|w\|_2^2 + m_0$ where $m_0 = \max_{i \in \mathcal{S}} \{V_i(x_0)\}$ and $\gamma = \max_{i \in \mathcal{S}} \{\gamma_i\}$.

Corollary 3 (Reliability) For $i \in \mathcal{S}$, let $\gamma_i > 0$ be given, and assume that there exist positive constants ξ_{ji} ($j = 1, 2, 3$), ϵ_i , α_i , control gain $K_i = -\frac{1}{2}\epsilon_i B_{i\bar{\Sigma}}^T P_i$, and a positive-definite matrix P_i such that the following Riccati-like equation holds

$$\begin{aligned}
&A_i^T P_i + P_i A_i + P_i (\xi_{1i} G_{ic} G_{ic}^T - \epsilon_i B_{i\bar{\Sigma}} B_{i\bar{\Sigma}}^T + \xi_{2i} \bar{A}_i (\bar{A}_i)^T + \xi_{3i} I) P_i \\
&+ \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I + C_{ic}^T C_{ic} + \alpha_i P_i = 0.
\end{aligned} \tag{17}$$

Assume further that $\|w_\sigma^F\|^2 \leq \xi_{1i} \alpha_i^* V_i(x)$, where $w_\sigma^F = (w^T \ (u_\sigma^F)^T)^T$ is the augmented disturbance input to the system, with $u_\sigma^F \in \mathbb{R}^q$ is the failure vector whose elements corresponding to the set of faulty actuators σ , $\alpha_i^* < \alpha_i$ and for all k , $r \leq t_k - t_{k-1} \leq \beta$ where $\beta > 0$, assume further that the ADT condition holds. Then, system (15) is globally exponentially ISS- H_∞ .

Proof For $i \in \mathcal{S}$, define the Lyapunov function candidate $V_i(x) = x^T P_i x$. Then

$$\begin{aligned}
\dot{V}_i(x) &\leq x^T [A_i^T P_i + P_i A_i + P_i (\xi_{1i} G_{ic} G_{ic}^T + \xi_{2i} \bar{A}_i (\bar{A}_i)^T - \epsilon_i B_{i\bar{\Sigma}} B_{i\bar{\Sigma}}^T \\
&\quad + \xi_{3i} I) P_i + \left(\frac{q_i}{\xi_{2i}} + \frac{\delta_i}{\xi_{3i}} \right) I] x + \frac{1}{\xi_{1i}} (w_\sigma^F)^T w_\sigma^F \leq -\alpha_i V_i(x) + \frac{1}{\xi_{1i}} (w_\sigma^F)^T w_\sigma^F \\
&\leq -\lambda_i V_i(x) \leq -\lambda V_i(x),
\end{aligned}$$

with $\lambda_i = \alpha_i - \alpha_i^*$ and $\lambda = \min_{i \in \mathcal{S}} \{\lambda_i\}$, where we have used Lemma 1, condition (17), and the fact that $B_{i\bar{\Sigma}} B_{i\Sigma}^T \leq B_{i\bar{\sigma}} B_{i\sigma}^T$ [21]. Thus, we have $\|x\| \leq b \|x_{t_0}\|_r e^{-\alpha(t-t_0)/2} + \bar{\gamma}(t)$, for $t \geq t_0$, where $b = \mu \sqrt{e^\eta}$, and $\bar{\gamma}(t) = \sqrt{\Gamma(t)/c_1}$ is class \mathcal{K} such that $\Gamma(s) = k(\mu^2)^{k-1} \frac{\|w_\sigma^F(s)\|^2}{\xi_1 \alpha^*}$ and $\xi_1 \alpha^* = \min_{i \in \mathcal{S}} \{\xi_{1i} \alpha_i^*\}$. Similarly, we can find the upper bound to $\|z\|$ where in this case $J_i = \int_{t_0}^\infty (z^T z - \gamma_i^2 (w_\sigma^F)^T w_\sigma^F) dt$.

4 Numerical Examples

Example 1 Consider system (14) with $\mathcal{S} = \{1, 2\}$. In the first mode, we have

$$A_1 = \begin{bmatrix} 0.2 & 0.1 \\ 0 & -6 \end{bmatrix}, B_1 = \begin{bmatrix} -3 & 1 \\ 0.1 & 0.2 \end{bmatrix}, C_1 = \begin{bmatrix} 2 & 0.1 \\ 0 & 2 \end{bmatrix}, F_1 = \begin{bmatrix} 0.1 & -2 \\ 0.1 & 0 \end{bmatrix},$$

$$\bar{A}_1 = \begin{bmatrix} 0.1 & 0.1 \\ 0.2 & 1 \end{bmatrix}, G_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, f_1 = 0.1 \begin{bmatrix} \sin(x_1(t)) \\ \sin(x_2(t)) \end{bmatrix},$$

and the tuning parameters: $\epsilon_1=2$, $\gamma_1 = 0.1$, $\alpha_1 = 2$, $\xi_{11} = \gamma_1^{-2}$, $\xi_{21} = 0.1$, $\xi_{31} = 0.2$, $M_1 = 2$, $\beta = 3$, $\theta_1 = 0.05$, and $\delta_1 = 0.1$. As for the second mode, we have

$$A_2 = \begin{bmatrix} -9 & 0.2 \\ 0 & 0.1 \end{bmatrix}, B_2 = \begin{bmatrix} 0.1 & 0.5 \\ 0.1 & -1 \end{bmatrix}, C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}, F_2 = \begin{bmatrix} 0.1 & 0 \\ -3 & 0.1 \end{bmatrix},$$

$$\bar{A}_2 = \begin{bmatrix} 0.3 & 0.2 \\ 0 & 0.1 \end{bmatrix}, G_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}, f_2 = 0.01 \begin{bmatrix} \sin(x_1(t)) \\ \sin(x_2(t)) \end{bmatrix},$$

and tuning parameters: $\epsilon_2=0.5$, $\gamma_2 = 0.15$, $\alpha_2=2.5$, $\xi_{12}=\gamma_2^{-2}$, $\xi_{22}=0.2$, $\xi_{32}=0.1$, $M_2 = 1.1$, $\theta_2 = 0.15$, and $\delta_2 = 0.01$. The disturbance is $w(t) = \begin{bmatrix} e^{-0.2t} \sin(t) \\ e^{-0.2t} \sin(t) \end{bmatrix}$,

Case 1 (Normal Actuators): From the Ricatti-like matrix equations, we obtain

$P_1 = \begin{bmatrix} 10.1452 & -1.0498 \\ -1.0498 & 9.2549 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 29.8698 & -2.2714 \\ -2.2714 & 15.3825 \end{bmatrix}$, and the control gains $K_1 = \begin{bmatrix} 30.5406 & -4.0748 \\ -9.9352 & -0.8012 \end{bmatrix}$ and $K_2 = \begin{bmatrix} -0.6900 & -0.3278 \\ -4.3016 & 4.1295 \end{bmatrix}$. So that, from condition (i) in Theorem 1, we get $c_1 = 8.5598$ and $c_2 = 30.2176$. Then, from the ADT condition, we get $\tau_a = \frac{\ln \mu}{\alpha^* - \nu} = 0.8699$ where $\nu = 0.5$, and the cheater bound is $N_0 = 0.5853$. The upper bound of the disturbance magnitude is 0.1031.

Case 2. When there is a failure in the first actuator, i.e., $B_{1\Sigma} = \{1\}$ and $B_{1\bar{\Sigma}} = \begin{bmatrix} 0 & 1 \\ 0 & 0.2 \end{bmatrix}$, and $B_{2\Sigma} = \{2\}$ and $B_{2\bar{\Sigma}} = \begin{bmatrix} 0.1 & 0 \\ 0.1 & 0 \end{bmatrix}$, then from the Ricatti-like equations,

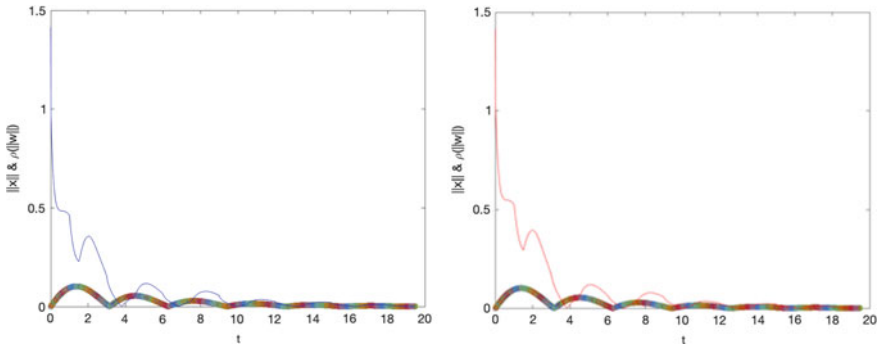


Fig. 1 ISS: operational case (left) and faulty case (right)

we obtain $P_1 = \begin{bmatrix} 10.2190 & -2.7934 \\ -2.7934 & 10.0408 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 29.6137 & -2.7372 \\ -2.7372 & 15.4077 \end{bmatrix}$, and the control gain matrices $K_1 = \begin{bmatrix} 30.9364 & -9.3843 \\ 0 & 0 \end{bmatrix}$ and $K_2 = \begin{bmatrix} 0 & 0 \\ -4.3860 & 4.1941 \end{bmatrix}$. From condition (i) in Theorem 1, we get $c_1 = 7.3351$, $c_2 = 30.1228$. Thus, $A_i + B_i K_i$ ($i = 1, 2$) are Hurwitz and the ADT is $\tau_a = 0.9742$ where the cheater bound is $N_0 = 0.5378$. The upper bound of the disturbance magnitude is 0.1033. Figure 1 shows the simulation results of $\|x\|$ (thin) and $\rho(s)$ (bold) for both cases, where $\rho(s) = \max\{\rho_1(s), \rho_2(s)\}$ and $\rho_i(s) = s/\sqrt{c_2\theta_i\xi_{2i}}$.

References

1. van der Schaft, A., Schumacher, H.: An Introduction to Hybrid Dynamical Systems, vol. 251, 1st edn. Springer-Verlag, London (2000)
2. Lygeros, J.: Lecture Notes on Hybrid Systems. Course Notes, p. 82 (2004)
3. Hespanha, J.P., Lecture Notes on Hybrid Control and Switched Systems, 2005
4. Morari, M., Thiele, L.: Hybrid Systems: Computational and Control (2005)
5. Alwan, M.S., Liu, X., Xie, W.C.: Comparison principle and stability of differential equations with piecewise constant arguments. *J. Franklin Inst.* **350**(2), 211–230 (2013)
6. Matveev, A.S., Savkin, A.V.: Qualitative Theory of Hybrid Dynamical Systems. Birkhauser, Cambridge, MA (2000)
7. Liberzon, D.: Switching in Systems and Control. Birkhäuser (2003)
8. Li, Z., Soh, Y., Wen, C.: Switched and impulsive systems: analysis, design and applications, vol. 313. Springer Science and Business Media (2005)
9. Pnueli, A., Sifakis, J.: Special Issue on Hybrid Systems of Theoretical Computer Science, vol. 138, no. 1. (1995)
10. Koutsoukos, X.D., He, K.X., Lemmon, M.D., Antsaklis, P.J.: Timed Petri nets in hybrid systems: stability and supervisory control. *Discrete Event Dyn. Syst.* **8**(2), 137–173 (1998)
11. Antsaklis, P.J., Nerode, A.: Hybrid control systems: an introductory discussion to the special issue. *IEEE Trans. Auto. Control* **43**(4), 457–460 (1998)
12. Antsaklis, P.J.: A brief introduction to the theory and applications of hybrid systems. *Proc. IEEE Special Issue Hybrid Syst. Theory Appl.* **88**(7), 879–887 (2000)

13. Hespanha, J.P., Morse, A.S.: Stability of switched systems with average dwell-time. In: Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, vol. 3, pp. 2655–2660. (1999)
14. Cheng, X.M., Gui, W.H., Gan, Z.J.: Robust reliable control for a class of time-varying uncertain impulsive systems. *J. Central South Univ. Technol.* **12**(1), 199–202 (2005)
15. Seo, C.J., Kim, B.K.: Robust and reliable H_∞ control for linear systems with parameter uncertainty and actuator failure. *Automatica* **32**(3), 465–467 (1996)
16. Veillette, R.J., Medanic, J.V., Perkins, W.R.: Design of reliable control systems. *IEEE Trans. Auto. Control* **37**(3), 290–304 (1992)
17. Wang, J., Shao, H.: Delay-dependent robust and reliable H_∞ control for uncertain time-delay systems with actuator failures. *J. Franklin Inst.* **337**, 781–791 (2000)
18. Yang, G.H., Wang, J.L., Soh, Y.C.: Reliable H_∞ control design for linear systems. *Automatica* **37**(5), 717–725 (2001)
19. Chen, G., Xiang, Z.: Robust reliable H_∞ control of switched stochastic systems with time delays under asynchronous switching. *Adv. Differ. Equ. Spring. Open J.* **86** (2013)
20. Alwan, M.S., Liu, X.Z., Xie, W.-C.: On design of robust reliable H_∞ control and input-to-state stabilization of uncertain stochastic systems with state delay. *Commun. Nonlinear Sci. Numer. Simul.* **18**(4), 1047–1056 (2013)
21. Seo, C.-J., Kim, B.K.: Robust and reliable H_∞ control for linear systems with parameter uncertainty and actuator failure. *Automatica* **32**(3), 3–5 (1996)
22. Sontag, E.D.: Smooth stabilization implies coprime factorization. *IEEE Trans. Automatic Control* **34**(4), 435–443 (1989)
23. Khalil, H.K.: *Nonlinear Systems*, 3rd edn. Prentice-Hall (2002)
24. Teel, A.R., Moreau, L., Nešić, D.: A note on the robustness of input-to-state stability. In: Proceeding of the 40th IEEE in decision and control, Florida, vol. 1, pp. 875–880. (2001)
25. Kuang, Y.: Delay differential equations with applications in population dynamics. *Mathematics in Science and Engineering*, vol. 191. Academic Press, INC. (1993)

Impulsive Distance-Based Formation Tracking Control of Multi-agent Systems



Zixing Wu, Xinzhi Liu, and Jinsheng Sun

Abstract In this paper, we discuss the distance-based formation control problems for double-integrator multi-agent system (MAS) via impulsive protocols. The proposed controller allows all agents to attain the desired formation shape by controlling the inter-distances and velocities. Unlike the common conditions in most literatures associated with distance-based formation, where the information is assumed to be exchanged continuously, in our proposed strategy, the information of leader are exchanged according to an impulsive sequence, which is more applicable to some difficult communication environment. For the continuous control term, we will generate a potential function and use relative distance information of inter-agents to form and keep the desired formation shape. The impulsive control term is applied to the tracking errors between each agent and the leader at every impulse moment. By using Lyapunov's method, the exponential stability of the system with the proposed impulsive control law has been demonstrated. Numerical examples are provided to validate the effectiveness of our approach.

Keywords Multi-agent system · Formation control · Impulsive control · Cooperative control

1 Introduction

Formation control of networked multi-agent systems has received considerable attention in recent years due to its extensive applications. The formation control can be

Z. Wu (✉) · J. Sun
Nanjing University of Science and Technology, Jiangsu, China
e-mail: wuzixing@njjust.edu.cn

J. Sun
e-mail: jssun67@163.com

X. Liu
University of Waterloo, Ontario, Canada
e-mail: xinzhi.liu@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_22

233

classified into position-based, displacement-based, bearing-based and distance-based formation control according to the controlled variable and sensed variable [1]. The distance-based formation control has recently attracted research interest, because it reduced requirement on the sensing capability for individual agents and do not share a common coordinate system [2].

With underpinnings from rigid graph, the distance-based formation control problem in continuous-time has been studied in the literature, see for example [3, 4]. However, most of the results so far are based on the assumption that all the information exchanges in the MAS are completed during the full response time. As is well known, continuous-time communication is always too expensive or unavailable. By contrast, for the impulse control method, the state information is just transmitted at impulsive instants, and the state information transmitted among the nodes of multi-agent systems is reduced greatly. Impulsive control has gained considerable interest in various areas. Recently, Ref. [5] designed an impulsive algorithms by using only the relative position to achieve the formation tracking. In [6], some results were presented on containment control in multi-agent systems that have static or dynamic leaders under directed and undirected communication topologies by periodic impulsive algorithms. Although much researches have been done on consensus problem [7, 8] and formation control [5, 9, 10] of MASs, this is to our best knowledge the first time to incorporate an impulsive protocol with distance-based formation control.

Inspired by the above discussion, a novel distributed formation control strategy for double-integrator multi-agent is introduced. The algorithms presented in this paper only need distance information of inter-agent and the communication graph and desired formation shape graph are assumed to be rigid. This is very significant because position or relative position according to the global coordination system can not be obtained in some situations. The impulsive protocols need the information of agents only at every impulsive instant, and regulate the velocities of all followers at every instant. Some necessary and sufficient conditions are also derived in this paper.

The overall arrangement of this paper is organized as follows. In Sect. 2, we demonstrate the preliminaries and problem formulation. The distance-based formation tracking control problem via impulsive protocol is studied in Sect. 3. Numerical example of dynamic leaders is presented to verify the validity of the proposed protocols in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Preliminaries and Problem Formulation

2.1 Graph Theory and Some Useful Lemmas

Consider an undirected graph with m edges and n vertices, denoted by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with vertex set $\mathcal{V}(\mathcal{G}) = \{v_1, \dots, v_n\}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The neighbor set of node i is defined as $\mathcal{N}_i := \{j \in \mathcal{V} : (i, j) \in \mathcal{E}\}$. The adjacency matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ is defined with nonnegative elements. The adjacency elements associated with

the edges of the graph are positive, i.e., $e_{ij} \in \mathcal{E} \Leftrightarrow a_{ij} > 0$. Moreover, we assume that $a_{ii} = 0$. The Laplacian matrix $L = [l_{ij}] \in \mathbb{R}^{n \times n}$ associated with the adjacency matrix A is defined by $l_{ij} = -a_{ij}$ for $i \neq j$, and $l_{ii} = \sum_{j=1, j \neq i}^N a_{ij}$ for $i = 1, 2, \dots, N$ which satisfies that $\sum_{j=1}^N l_{ij} = 0$. For a connected graph \mathcal{G} , L has a simple zero eigenvalue and all the other eigenvalues have positive real parts. All these eigenvalues can be ordered as $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$.

In this paper, we consider a rigid formation modeled by an undirected graph. Then, a matrix relating to the nodes to the edges is introduced, which is called the incidence matrix $H = [h_{ij}] \in \mathbb{R}^{n \times m}$ whose entries are defined as

$$h_{ij} = \begin{cases} 1, & \text{the } i \text{ th edge sinks at node } j, \\ -1, & \text{the } i \text{ th edge leaves node } j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that for a rigid formation modeled by an undirected graph considered in this paper, the orientation of each edge for writing the incidence matrix can be defined arbitrarily.

Let $p_i \in \mathbb{R}^d$ where $d = 2, 3$ denote a point that is assigned to $i \in \mathcal{V}$. The stacked vector $p = [p_1^T, p_2^T, \dots, p_n^T]^T \in \mathbb{R}^{dn}$ represents the realization of \mathcal{G} in \mathbb{R}^d . The rigidity function $r_{\mathcal{G}}(p) : \mathbb{R}^{dn} \rightarrow \mathbb{R}^m$ associated with the framework (\mathcal{G}, p) is given as:

$$r_{\mathcal{G}}(p) = \frac{1}{2} \left[\dots, \|p_i - p_j\|^2, \dots \right]^T, (i, j) \in \mathcal{E},$$

where $\|\cdot\|$ is the standard Euclidean norm, the k -th component in $r_{\mathcal{G}}(p)$, $\|p_i - p_j\|^2$ corresponds to the squared length of the relative position vector z_k which connected vertices i and j . One useful tool to characterize the rigidity property of a framework is the rigidity matrix $R \in \mathbb{R}^{m \times dn}$, which is defined as

$$R(p) = \frac{\partial r_{\mathcal{G}}(p)}{\partial p}. \tag{1}$$

The row of the rigidity matrix R corresponding to $\{(i, j) \in \mathcal{E}\}$ takes the following form

$$\left[\mathbf{0}_{1 \times d}, \dots, (p_i - p_j)^T, \dots, \mathbf{0}_{1 \times d}, \dots, (p_j - p_i)^T, \dots, \mathbf{0}_{1 \times d} \right].$$

The following Lemmas are needed for later use.

Lemma 1 [11] *Let X and Y be arbitrary n -dimensional real vectors, $P \in \mathbb{R}^{n \times n}$ be a positive definite matrix, and $\varepsilon > 0$. Then, the following matrix inequality holds:*

$$\pm \{X^T P^T Y + Y^T P X\} \leq \varepsilon X^T P X + \varepsilon^{-1} Y^T P Y. \tag{2}$$

2.2 Problem Statement

Consider the following n double-integrator agents in d -dimensional space.

$$\begin{cases} \dot{p}_i(t) = q_i(t) \\ \dot{q}_i(t) = u_i(t) \end{cases}, i = 1, \dots, n, \quad (3)$$

where $p_i \in \mathbb{R}^d$, $q_i \in \mathbb{R}^d$, $u_i \in \mathbb{R}^d$, $d = \{2, 3\}$ denote the position, the velocity and the control input for $i \in \mathcal{V}$. The leader dynamic is expressed as

$$\dot{p}_r(t) = q_r(t), \quad \dot{q}_r = u_r(p_r, t), \quad (4)$$

where p_r , q_r and $u_r(p_r, t)$ denote the position, velocity and accelerated velocity of leader. $(p_r(0), q_r(0))$ and $u_r(p_r, t)$ are given in advance as a reference.

In distance-based formation setup, the desired formation shape is described with distance constraints. Given a formation shape realization $p^* = [p_1^T, p_2^T, \dots, p_n^T]^T \in \mathbb{R}^{dn}$. We define the desired distance for the k -th edge ($k = 1, 2, \dots, m$) which connects the agent i and j as $d_{ij} = \|p_i^* - p_j^*\|$ and the distance error is defined as

$$e_k = \|p_i - p_j\|^2 - d_{ij}^2. \quad (5)$$

Denote the error vector as $e = [e_1, e_2, \dots, e_m]^T \in \mathbb{R}^m$. Further, the desired formation E_{p^*, v^*} of the agents can be expressed as

$$E_{p^*, q^*} := \{[p^T, q^T]^T \in \mathbb{R}^{2dn} : e = 0, q_i = q_r\}. \quad (6)$$

Then the impulsive distance formation control problem for the double-integrator modeled agents is stated as follows:

Consider a group of n double-integrator modeled agents (3) in d -dimensional space, suppose that the sensing graph of the agents is given by an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Design a distributed controller u_i with given the impulsive sequence, such that the distances between each agent reach defined distance constraints and agents maintain consistent velocity with the leader, which means that E_{p^*, q^*} is achieved and the system (3) with controller is exponential stable under the impulsive distributed control law.

3 Main Results

3.1 Undirected Formations of Double Integrators

The formation control law is designed as:

$$\begin{aligned}
 u_i(t) = & u_r(t) - \sum_{j \in \mathcal{N}_i} (p_i(t) - p_j(t)) (\|p_i(t) - p_j(t)\|^2 - d_{ij}^2) \\
 & - k_v \sum_{j \in \mathcal{N}_i} a_{ij} (q_i(t) - q_j(t)) \\
 & - \sum_{k=1}^{\infty} [c_i (p_i(t_k) - p_r(t_k)) + k_v c_i (q_i(t_k) - q_r(t_k))] \delta(t - t_k),
 \end{aligned} \tag{7}$$

where $k_v > 0$, $c_i \geq 0$ are constant non-negative coefficients, $c_i = 0$ means that the i -th agent cannot receive the information from the reference, a_{ij} is the element of adjacency matrix A . Define $C = \text{diag}\{c_i\} \in \mathbb{R}^{n \times n}$, then the eigenvalues of matrix C (denoted by $\lambda(C)$) depend on the choice of elements c_i and one has $\lambda_{\max}(C) = \sup\{c_i\}$. Dirac function $\delta(t)$ denotes the impulsive effects at the time moment $t = t_k$, the time sequence $\{t_k\}$ with $0 = t_0 < t_1 < \dots < t_k < \dots$, $k = 1, 2, \dots$, $\lim_{t \rightarrow \infty} t_k = \infty$ forms a strictly increasing sequence in the time interval $[0, \infty)$. The impulsive distances are defined as $\Delta t_k = t_k - t_{k-1}$.

Define $x_i = p_i - p_r$, $v_i = q_i - q_r$. The double-integrator model (3) combined with the controller (7) can be transformed into

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = -k_p \sum_{j \in \mathcal{N}_i} (x_i(t) - x_j(t)) (\|x_i(t) - x_j(t)\|^2 - d_{ij}^{*2}) \\ \quad - k_v \sum_{j \in \mathcal{N}_i} a_{ij} (q_i(t) - q_j(t)), & t \in [t_{k-1}, t_k), \\ \Delta v_i(t_k) = -c_i x_i(t_k) - c_i k_v v_i(t_k), & t = t_k, \end{cases} \tag{8}$$

which can be rewritten to a matrix form as

$$\begin{cases} \dot{x}(t) = v(t), \\ \dot{v}(t) = -R^T(x(t))e(x(t)) - k_v(L \otimes I_d)v(t), & t \in [t_k, t_{k+1}), \\ \Delta v(t_k) = -C(x(t_k) + k_v v(t_k)), & t = t_k, \end{cases} \tag{9}$$

where $R(x(t)) = R(p)$ is the rigid matrix and $\hat{L} = L \otimes I_d$. To deal with the position system with the impulsive protocol (9), we analyze the distance error system. By noting that $\dot{e}(t) = 2R(x)\dot{p}(t)$, one can obtain the following equation for the distance error system with the controller (7):

$$\dot{e}(t) = 2R(x)v(t), \quad t \in [t_{k-1}, t_k). \tag{10}$$

Then, based on the control inputs, the stability properties for the distance-based formation problem for double-integrator agents are stated in the following theorem:

Theorem 1 *Suppose the target formation is infinitesimally and minimally rigid. The error system (9) is exponentially stable if there exist positive scalars $\varepsilon > 0$, $0 < \mu \leq \rho < 1$ such that*

1.

$$\begin{pmatrix} (\varepsilon^{-1} - \alpha)I & 0 \\ * & (\varepsilon - \alpha)I - 2k_v \hat{L} \end{pmatrix} < 0,$$

2. *There exist constant $0 < \mu < 1$ satisfying*

$$\begin{pmatrix} C^T C + (\frac{1}{2} - \mu)I & -C^T(I - k_v C) \\ * & (I - k_v C)^T(I - k_v C) - \mu I \end{pmatrix} < 0,$$

3.

$$\ln \rho - \alpha h_2 < 0.$$

where $0 < h_1 < t_k - t_{k-1} < h_2$ is the impulsive period.

Proof Consider the following Lyapunov function candidate:

$$V = \frac{1}{2}e^T(t)e(t) + x^T(t)x(t) + v^T(t)v(t). \quad (11)$$

For $t \in [t_{k-1}, t_k]$, $k \geq 0$, calculate the Dini's derivative of $V(t)$ as

$$\begin{aligned} D^+V(t) &= e^T(t)P\dot{e}(t) + 2x^T(t)v(t) + 2v^T(t)\dot{v}(t) \\ &= 2e^T(t)R(x)v(t) + 2x^T(t)v(t) - 2v^T(t)R(x)e(t) - 2k_v v^T(t)\hat{L}v(t) \\ &\leq \varepsilon^{-1}x^T(t)x(t) + \varepsilon v^T(t)v(t) - 2k_v v^T(t)\hat{L}v(t), \end{aligned} \quad (12)$$

From condition 1, we have

$$D^+V(t) \leq \alpha V(t),$$

which yields to $V(t) \leq e^{\alpha(t-t_k)} V(t_k^+)$, $t \in [t_{k-1}, t_k]$.

For $t = t_k$, we have $x(t_k^+) = x(t_k^-)$, $e(t_k^+) = e(t_k^-)$. There holds

$$\begin{aligned} v(t_k^+) &= \frac{1}{2}e^T(t_k^-)e(t_k^-) + x^T(t_k^-)x(t_k^-) \\ &\quad + \left(-Cx(t_k^-) + (I - k_v C)v^T(t_k^-)\right)^T \left(-Cx(t_k^-) + (I - k_v C)v(t_k^-)\right) \\ &\leq \frac{1}{2}e^T(t_k^-)e(t_k^-) + \begin{pmatrix} x(t_k^-) \\ v(t_k^-) \end{pmatrix}^T \begin{pmatrix} C^T C + \frac{1}{2}I & -C^T(I - k_v C) \\ * & (I - k_v C)^T(I - k_v C) \end{pmatrix} \begin{pmatrix} x(t_k^-) \\ v(t_k^-) \end{pmatrix}, \end{aligned}$$

By condition 2, we have $V(t_k^+) \leq \frac{1}{4}e^T(t_k^-)e(t_k^-) + \frac{1}{2}\mu x^T(t_k^-)x(t_k^-) + \frac{1}{2}\mu x^T(t_k^-)x(t_k^-)$. It follows that $V(t)$ is decreasing at every impulsive instant. Consequently, there exist positive factor $0 < \mu \leq \rho < 1$, such that

$$V(t_k^+) \leq \rho V(t_k^-).$$

It can be obtained that

$$V(t) \leq \rho^{k-1} e^{\alpha(t-t_0)} V(t_0), \quad t \in [t_{k-1}, t_k).$$

For $t \in [t_0, t_1)$

$$V(t) \leq e^{\alpha(t-t_0)} V(t_0).$$

For $t \in [t_1, t_2)$

$$V(t_1^+) \leq \rho e^{\alpha(t_1-t_0)} V(t_0).$$

Thus we conclude from mathematical induction that, for $t \in [t_k, t_{k+1})$,

$$V(t) \leq \rho^k e^{\alpha(t-t_0)} V(t_0). \quad (13)$$

Hence, as $t \rightarrow \infty$ the system (3) will converge to the set E_{p^*, q^*} where the distance error $\lim_{t \rightarrow \infty} e(t) = 0$ means that the desired formation construction is achieved and the tracking errors $\lim_{t \rightarrow \infty} x(t) = 0$, $\lim_{t \rightarrow \infty} v(t) = 0$ means that agents follow the movement of the reference.

This completes the proof. ■

Remark 1 It should be pointed out that in contrast to the common impulsive control systems, the distributed control input designed in this paper is not zero when $t \neq t_k$, such that the agents can receive the inter-agent relative position information. In fact, the impulsive effects are only associated with the information of the leader.

Remark 2 The distance error for the k -th edge $e_k = \|p_i - p_j\|^2 - d_{ij}^2$ is related to the relative position of the agent i and j . Since the control objective is to achieve desired formation shape which is restricted by relative distance, we conclude that the distance-based formation can be obtained when $\lim_{t \rightarrow \infty} e(t) = 0$.

4 Numerical examples

In this section, an illustrative example is provided to verify the effectiveness of the proposed impulsive distance-based formation control method for five double-integrator agents in a plane.

Example 1 The desired formation was set to the regular convex pentagon with the desired distances of connected edges were $d_{12} = d_{15} = d_{23} = d_{34} = d_{45} =$

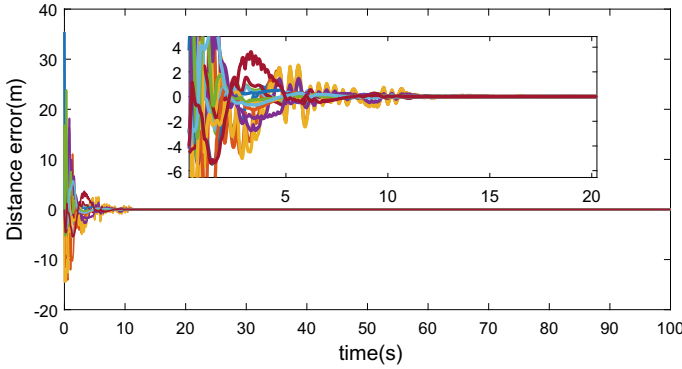


Fig. 1 The evolutions of distance errors with desired formation shape

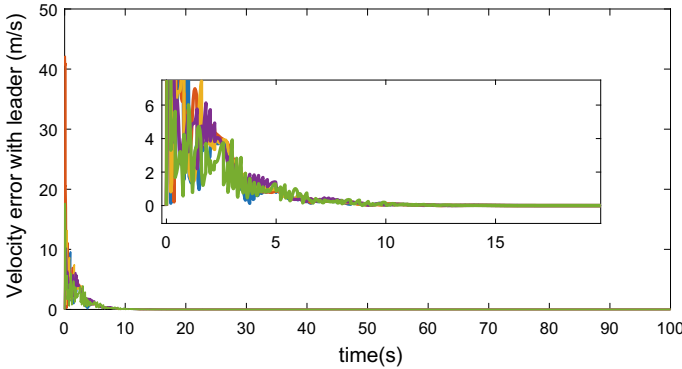
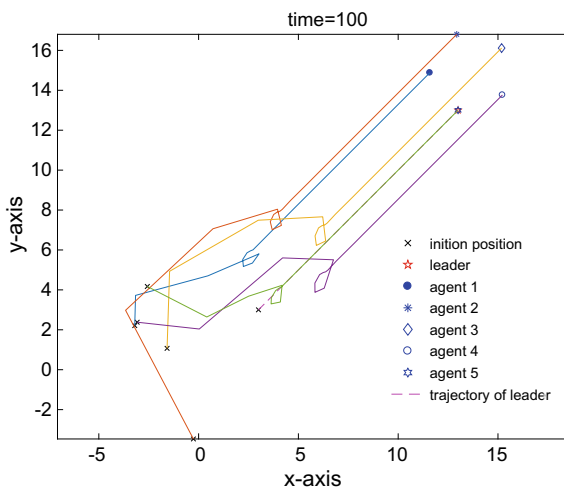


Fig. 2 The evolutions of velocity errors with the leader

$\sqrt{2(1 - \cos 2\pi/5)}$ and $d_{13} = d_{14} = \sqrt{2(1 + \cos \pi/5)}$. The control impulsive interval is set as a constant $0.2s$ in the simulation. The components of the initial positions $p_i(0)$ for each agent are randomly perturbed from those of p_i^* by a random variable uniformly distributed on $[-5,5]$, and the initial velocities for each agent are set as zeros. The motion of the leader is set as $u_r = 0$ and the initial velocity of leader is $[0.5, 0.5]^T$. The parameters in controller are set as $k_v = 0.2$ and $C = \text{diag}\{0, 0, 0, 0, 0.8\}$. Figs. 1 and 2 illustrate distance errors and the norm of velocity errors with the leader convergence to zeros. The trajectories of each agent, together with the initial position and the final shape are depicted in Fig. 3.

Remark 3 The leader can be a virtual reference or a real agent with a given motion. Since we only focused on the tracking of the velocity state of the leader, the controller only contains the information of the velocity of leader. As shown in Fig. 2, the norm of velocity errors converge to zero. In the simulations, the step size $0.001s$ are used. The control impulsive interval is chosen as $\Delta_{t_k} = 0.2s$.

Fig. 3 The trajectories of five followers and one leader. Only the agent 5 can receive the information from leader. The desired formation shape is achieved and the whole formation moving with the leader



5 Conclusions

In this paper, leader-following distance-based formation control of double-integrator MASs under impulsive protocol is investigated. The impulsive formation tracking algorithms with the leader information have been proposed to achieve the formation. Based on the stability theory of impulsive systems and rigid graph theory, the exponentially stability for the closed-loop system combined with the system and the hybrid controller has been proved. Finally, an example has been given to illustrate the effectiveness of our theoretical results.

References

1. Oh, K.K., Park, M.C., Ahn, H.S.: A survey of multi-agent formation control. *Automatica* **53**, 424–440 (2015)
2. Sun, Z., Anderson, B.D., Deghat, M., Ahn, H.S.: Rigid formation control of double-integrator systems. *Int. J. Control* **90**(7), 1403–19 (2016)
3. Sun, Z., Anderson, B.D., Deghat, M., Ahn, H.S.: Rigid formation control of double-integrator systems. *Int. J. Control* **90**, 1403–1419 (2016)
4. Colombo, L.J., de Marina, H.G.: A Variational Integrator for the distance-based formation control of multi-agent systems. *IFAC-PapersOnLine* **51**(23), 76–81 (2018)
5. Wang, Y.W., Liu, M., Liu, Z.W., Yi, J.W.: Formation tracking of the second-order multi-agent systems using position-only information via impulsive control with input delays. *Appl. Math. Comput.* **246**, 572–585 (2014)
6. Zhang, H.X., Ding, L., Liu, Z.W.: Schooling for multi-agent systems via impulsive containment control algorithms with quantized information. *Trans. Inst. Measur. Control* **41**, 828–841 (2018)
7. Liu, X., Zhang, K., Xie, W.C.: Impulsive consensus of networked multi-agent systems with distributed delays in agent dynamics and impulsive protocols. *J. Dyn. Syst. Measur. Control* **141**, 011008-08-8 (2018)

8. Zhu, W., Wang, D.: Leader-following consensus of multi-agent systems via event-based impulsive control. *Measur. Control* **52**(1–2), 91–99 (2019)
9. Gaias, G., D’Amico, S.: Impulsive maneuvers for formation reconfiguration using relative orbital elements. *J. Guidance Control Dyn.* **38**, 1036–1049 (2015)
10. Qin, W., Liu, Z., Chen, Z.: Impulsive formation control algorithms for leader-following second-order nonlinear multi-agent systems. *IFAC Proc. Vol.* **46**, 172–77 (2013)
11. Wu, J., Jiao, L.: Synchronization in complex delayed dynamical networks with nonsymmetric coupling. *Physica A* **386**, 513–C530 (2007)

Exponential Stabilization for Markov Jump Neural Networks with Additive Time-Varying Delays via Event-Triggered Impulsive Control



Haiyang Zhang, Zhipeng Qiu, Xinzhi Liu, and Lianglin Xiong

Abstract This paper investigates the Exponential Stabilization (ES) problem for Markov Jumping Neural Networks (MJNNs) with Additive Time-varying Delays (ATDs). To further mitigate the “unnecessary” waste of networks resources, a Sample-based Event-triggered Impulsive Control (SEIC) scheme is employed. A novel Lyapunov-Krasovskii functional is constructed by considering more information about sampled data, ATDs and Markov jump parameters. In virtue of the SEIC scheme, a new ES criterion for MJNNs with ATDs is then presented. In the end, a numerical example is given to illustrate the validity of the obtained result.

Keywords Exponential stabilization · Neural networks · Markov jump parameters · Additive time-varying delays · Event-triggered impulsive control

1 Introduction

In recent decades, Neural Networks (NNs) have received considerable attentions since its extensive applications in many different fields, such as pattern recognition [1–3], smart antenna arrays [4–6], and so forth. Time-varying Delays (TDs) are inevitably encountered in NNs [7] due to the inherent communication time among the

H. Zhang (✉) · L. Xiong
School of Mathematics and Computer Science, Yunnan Minzu University,
Kunming 650500, China
e-mail: haiya287@126.com

L. Xiong
e-mail: lianglin_5318@126.com

Z. Qiu
School of Science, Nanjing University of Science and Technology, Nanjing 210094, China
e-mail: nustqzp@njust.edu.cn

X. Liu
Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: xzliu@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_23

243

neurons, and other reasons. Especially, when signals are transferred from one node to another, Additive Time-varying Delays (ATDs) with different physical characteristics are produced in NNs, because the transmission channel and circumstances may be entirely distinct in different segments of networks. As a result, some conservativeness could be generated if the different kinds of ATDs are regarded as the same [8]. It is known that stability is a precondition for the implementation of systems, but the existence of TDs often leads to NNs chaotic, oscillation and even unstable [9]. In addition, fast convergence of the networks is essential for realtime computation, and the exponential-convergence rate is generally used to determine the speed of neural computations [10]. Thus, it is of great theoretical and practical importance to study the exponential stability for NNs with ATDs.

As is well known, the structures and parameters of NNs are often subjected to random abrupt variations [11], such as external environment sudden change, information latching, and so on. Markov Jump systems (MJSs), as a special kind of hybrid systems, have a powerful ability to describe those random behaviors [12]. In addition, Impulsive control, as a powerful tool, plays an important role in many different science and engineering fields. Especially, in the field of artificial neural networks, the study for stabilization is more complicated due to the state-dependent nonlinear switching behaviors of NNs. Consequently, the research about exponential stabilization of Impulsive Markov jump neural networks (IMJNNs) with TDs have attracted much more attention [13–18]. However, there are few works about exponential stabilization for IMJNNs with ATDs, which leaves much room for investigation.

On the other hand, in traditional impulsive control strategy [13–15], the impulse signals are transmitted periodically, or the impulsive instants are predesigned. As a result, some “unnecessary” data are sent frequently and the network resources are excessively used. It is necessary to improve the traditional control scheme, especially in the case that the networks resources are limited [19]. Therefore, to reduce unnecessary waste of network resources, an alternative control scheme, namely, event-triggered control scheme was proposed [20], and then Event-triggered Impulsive Control (EIC) scheme was developed in [21–25]. It should be pointed out that the above EIC scheme is in the means of continuous-time, which has some disadvantages. For example, it requires sensors to monitor the system state all the time, but it is not necessary due to the worst scenario rarely happens. Thus, in order to further save the network resources, a SEIC scheme is adopted to investigate the exponential stabilization for IMJNNs with ATDs in this paper.

Notations: Let \mathbb{N} denote the set of positive integers, \mathbb{R} the set of real numbers, and \mathbb{R}^n the n -dimensional real space equipped with the Euclidean norm $\|\cdot\|$, $\mathbb{R}^{m \times n}$ the set of all $m \times n$ real matrices, \mathbb{S}_+^n and \mathbb{S}^n the set of symmetric positive definite and symmetric matrices of $\mathbb{R}^{n \times n}$, respectively. The symbol “*” in a block matrix signifies the symmetric terms, $col\{\cdot\cdot\cdot\}$ and $diag\{\cdot\cdot\cdot\}$ express a column vector and a diagonal matrix, respectively. For any matrix $X \in \mathbb{R}^{n \times n}$, $\mathbb{H}\{X\}$ means that $X + X^T$, $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ stand for the maximum and minimum eigenvalue of X , respectively. The zero and identity matrices with appropriate dimensions are described by 0 and I , respectively.

2 Description of problem and preliminaries

Let $\{r(t), t \geq 0\}$ be a continuous-time Markov Processes (MPs) taking values in a finite state space $\mathfrak{N} = (1, 2, \dots, N)$. The evolution of MPs $\{r(t), t \geq 0\}$ is governed by the following transition probability:

$$Pr\{r(t + \Delta) = j \mid r(t) = i\} = \begin{cases} \pi_{ij}\Delta + o(\Delta), & i \neq j, \\ 1 + \pi_{ii}\Delta + o(\Delta), & i = j, \end{cases} \quad (1)$$

where $\Delta \geq 0$, $\lim_{\Delta \rightarrow 0} o(\Delta)/\Delta = 0$; $\pi_{ij} \geq 0$ for $i \neq j \in \mathfrak{N}$ is the Transition Rate (TR) from mode i at time t to mode j at time $t + \Delta$, and $\pi_{ii} = -\sum_{j=1, j \neq i}^N \pi_{ij}$.

Consider the following impulsive MJNNs with ATDs:

$$\begin{cases} \dot{x}(t) = -B_{r(t)}x(t) + A_{r(t)}f(x(t)) + C_{r(t)}x(t_k) \\ \quad + D_{r(t)}f(x(t - \delta_1(t) - \delta_2(t))), \quad t \neq t_k, \\ x(t^+) = (1 + q_k)x(t^-), \quad t = t_k, \quad k \in \mathbb{N}, \\ \phi(\theta) = x(t_0 + \theta), \quad r(0) = r_0, \quad \theta \in [-\max\{\delta_1 + \delta_2, \eta\}, 0], \end{cases} \quad (2)$$

where $x(t) = \text{col}\{x_1(t), \dots, x_n(t)\}$ is the state; $f(x(t)) = \text{col}\{f_1(x_1(t)), \dots, f_n(x_n(t))\}$ is the neuron activation function, and satisfies :

$$\lambda_l^- \leq \frac{f_l(y_1) - f_l(y_2)}{y_1 - y_2} \leq \lambda_l^+, \quad f_l(0) = 0, \quad l \in \mathbb{N}, \quad \forall y_1 \neq y_2 \in \mathbb{R}, \quad (3)$$

where λ_l^-, λ_l^+ are scalars which can be positive, negative and zero; $\phi(\theta)$ is the initial condition; $\delta_1(t), \delta_2(t)$ are two ATDs with different physical property, and satisfy:

$$0 \leq \delta_1(t) \leq \delta_1, \quad \dot{\delta}_1(t) \leq \mu_1 < 1, \quad 0 \leq \delta_2(t) \leq \delta_2, \quad \dot{\delta}_2(t) \leq \mu_2 < 1, \quad (4)$$

where δ_1, δ_2 , and μ_1, μ_2 are known constants; $B_{r(t)}$ is a positive diagonal matrix, $A_{r(t)}, D_{r(t)}$ are connection weighted matrices; $C_{r(t)}$ is a control gain matrix to be determined, q_k is the impulsive intensity, $r_0 \in \mathfrak{N}$ is the initial mode as $t = 0$.

To mitigate unnecessary waste of network resources, a Sample-based Event-triggered Impulsive Control (SEIC) scheme is employed in this paper. Assume that the system's state is periodically sampled, and the sampling sequence is depicted by the set $\Pi_s = \{0, h, 2h, \dots, kh\}$ with $k \in \mathbb{N}$, where h is a constant sampling period, and the event-triggered sequence is described by the set $\Pi_e = \{0, b_1h, b_2h, \dots, b_kh\} \subseteq \Pi_s$ with $b_k \in \mathbb{N}$. Suppose the event-triggered instants to be the impulsive instants, i.e., $t_k = b_kh$, then the next impulsive instant $t_{k+1} = t_k + l_mh$, where

$$l_m = \min\{l \mid |e^T(t_k + lh)\Omega e(t_k + lh)| > \sigma x^T(t_k)\Omega x(t_k)\}, \quad l \in \mathbb{N}, \quad (5)$$

and $\sigma \in [0, 1)$ is a constant trigger threshold, $\Omega \in \mathbb{S}_+^n$ is an unknown weighted matrix, $e(t_k + lh) = x(t_k + lh) - x(t_k)$ expresses the error between the two states at the latest trigger instant and the current sampling one.

For the sake of introducing the SEIC scheme to determine whether the current sampled-data should be transmitted, an effective way is to consider the sampled-data error at every sampling instant. Decompose the impulsive interval $[t_k, t_{k+1})$ into the following subintervals: $[t_k, t_{k+1}) = \bigcup_{l=0}^{m-1} \mathbb{I}_k(l)$, where $\mathbb{I}_k(l) = [t_k + lh, t_k + (l + 1)h)$. Define a function

$$\eta(t) = t - (t_k + lh), \quad t \in \mathbb{I}_k(l). \tag{6}$$

Note that $\eta(t)$ is a linear piecewise function and satisfies $0 \leq \eta(t) \leq h$, $\dot{\eta}(t) = 1, \forall t \in \mathbb{I}_k(l)$. For simplifying some notations, denote $\delta(t) = \delta_1(t) + \delta_2(t)$, $\delta = \delta_1 + \delta_2$, $\mu = \mu_1 + \mu_2$, and $B_i = B_{r(t)}$, $A_i = A_{r(t)}$, $D_i = D_{r(t)}$, $C_i = C_{r(t)}$ when $r(t) = i$. Then, combining with (5) and (6), the system (2) can be rewritten as

$$\begin{cases} \dot{x}(t) = -B_i x(t) + A_i f(x(t)) + D_i f(x(t - \delta(t))), \\ \quad + C_i x(t - \eta(t)) - C_i e(t - \eta(t)), \quad t \neq t_k, \\ x(t^+) = (1 + q_k)x(t^-), \quad t = t_k, \quad k \in \mathbb{N}, \\ \phi(\theta) = x(t_0 + \theta), \quad r(0) = r_0, \quad \theta \in [-\max\{\delta, \eta\}, 0]. \end{cases} \tag{7}$$

The following definition and lemmas will be recalled, and play a key role to demonstrate our main result.

Definition 1 [26] The system (7) is said to be stochastically exponentially stable in the mean square sense with convergence rate $\alpha > 0$, if there exist a constant $M > 0$ for $\forall t \geq t_0$ such that

$$\mathbb{E} \{ \|x(t)\|^2 \} \leq M e^{-\alpha(t-t_0)} \mathbb{E} \left\{ \sup_{\theta \in [-\max\{\delta, \eta\}, 0]} \{ \|\phi(\theta)\|^2, \|\dot{\phi}(\theta)\|^2 \} \right\}. \tag{8}$$

Lemma 1 [27] For a matrix $R \in \mathbb{S}_+^n$, scalars a and b with $a < b$, a differentiable vector function $x(s) : [a, b] \rightarrow \mathbb{R}^n$, the following inequality hold

$$(b - a) \int_a^b x^T(s) R x(s) ds \geq \left[\int_a^b x(s) ds \right]^T R \left[\int_a^b x(s) ds \right]. \tag{9}$$

Lemma 2 [28] For scalars $\alpha_1, \alpha_2 \in (0, 1)$ satisfying $\alpha_1 + \alpha_2 = 1$, and matrices $R_1, R_2 \in \mathbb{S}_+^n, Y \in \mathbb{R}^{n \times n}$, the following inequality hold

$$\text{diag} \left\{ \frac{1}{\alpha_1} R_1, \frac{1}{\alpha_2} R_2 \right\} \geq \begin{pmatrix} R_1 & Y \\ * & R_2 \end{pmatrix}, \text{ if } \begin{pmatrix} R_1 & Y \\ * & R_2 \end{pmatrix} > 0. \tag{10}$$

3 Main Results

In this section, our purpose is to establish a new stochastic exponential stabilization condition for systems (7) via the SEIC scheme (5). Before presenting the main result, the following vector and functional are defined for convenience:

$$\begin{aligned} \xi(t) &= \text{col}\{x(t), f(x(t)), x(t - \delta(t)), f(x(t - \delta(t))), x(t - \delta), x(t - \eta(t)), \\ &\quad x(t - \eta), e(t - \eta(t)), x(t - \delta_1(t)), x(t - \delta_1(t) - \delta_2), \dot{x}(t)\}, \\ e_i &= \text{col}\{0, \dots, 0, \underbrace{I}_v, 0, \dots, 0\}, (v = 1, 2, \dots, 11). \end{aligned}$$

and consider the following stochastic Lyapunov-Krasovskii Functional (LKF):

$$V(x(t), r(t)) = V_1(x(t), r(t)) + V_2(x(t), r(t)) + V_3(x(t), r(t)), \tag{11}$$

where

$$V_1(x(t), r(t)) = x^T(t)P(r_t)x(t) + x^T(t_k)Sx(t_k), \tag{12}$$

$$\begin{aligned} V_2(x(t), r(t)) &= \int_{t-\delta(t)}^t e^{\alpha(s-t)} \left[x^T(s)Q_1x(s) + f^T(x(s))Q_2f(x(s)) \right] ds \\ &\quad + \int_{t-\delta}^t e^{\alpha(s-t)} x^T(s)Q_3x(s)ds + \int_{t-\eta}^t e^{\alpha(s-t)} x^T(s)Q_4x(s)ds \\ &\quad + \int_{t-\delta_1(t)}^t e^{\alpha(s-t)} x^T(s)Q_5x(s)ds + \int_{t-\delta_1(t)-\delta_2}^t e^{\alpha(s-t)} x^T(s)Q_6x(s)ds, \end{aligned} \tag{13}$$

$$\begin{aligned} V_3(x(t), r(t)) &= \delta_1\delta_2 \int_{t-\delta}^t \int_u^t e^{\alpha(s-t)} \dot{x}^T(s)R_1\dot{x}(s)dsdu \\ &\quad + \eta \int_{t-\eta}^t \int_u^t e^{\alpha(s-t)} \dot{x}^T(s)R_2\dot{x}(s)dsdu. \end{aligned} \tag{14}$$

Theorem 1 For given positive scalars $\delta_1, \delta_2, \mu_1, \mu_2, \eta, \alpha$ and σ , the system (7) is said to be stochastically exponentially stable in the mean square sense, if there exist matrices $P_i, S, Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, R_1, R_2 \in \mathbb{S}_+^n, X_1, X_2, N_i, K_i \in \mathbb{R}^{n \times n}$ and diagonal matrix $M_1, M_2 \in \mathbb{S}_+^n$ such that

$$(1 + q_k)^2 \lambda_{\max}(P_i + S) - \lambda_{\min}(P_i) > 0, k \in \mathbb{N}, \tag{15}$$

$$\inf \{t_{k+1} - t_k\} = \beta > \frac{\ln q}{\alpha}, k \in \mathbb{N}, \tag{16}$$

$$\mathcal{R}_1 = \begin{bmatrix} R_1 & * \\ X_1^T & R_1 \end{bmatrix} > 0, \mathcal{R}_2 = \begin{bmatrix} R_2 & * \\ X_2^T & R_2 \end{bmatrix} > 0, \tag{17}$$

$$\Phi_i = \Phi_{1i} + \Phi_{2i} + \Phi_{3i} + \Phi_{4i} + \Phi_{5i} < 0, \tag{18}$$

where $q = (1 + q_k)^{2\lambda_{\max}(P_i + S)} / \lambda_{\min}(P_i)$, $\mathcal{R} = \text{diag}\{\delta_2 \mathcal{R}_1, \delta_1 \mathcal{R}_1\}$ and

$$\begin{aligned} \Phi_{1i} &= \mathbb{H} \left\{ -e_1^T P_i e_{11} \right\} + e_1^T \left(\alpha P_i + \Pi(P_j) \right) e_1 + \alpha (e_6 - e_8)^T S (e_6 - e_8), \\ \Phi_{2i} &= e_1^T (Q_1 + Q_3 + Q_4 + Q_5 + Q_6) e_1 + e_2^T Q_2 e_2 - e^{-\alpha\delta} (1 - \mu) e_3^T Q_1 e_3 \\ &\quad - e^{-\alpha\delta} (1 - \mu) e_4^T Q_2 e_4 - e^{-\alpha\delta} e_5^T Q_3 e_5 - e^{-\alpha\eta} e_7^T Q_4 e_7 \\ &\quad - e^{-\alpha\delta_1} (1 - \mu_1) e_9^T Q_5 e_9 - e^{-\alpha\delta} (1 - \mu_1) e_{10}^T Q_6 e_{10}, \\ \Phi_{3i} &= e_{11}^T (\delta\delta_1 \delta_2 R_1 + \eta^2 R_2) e_{11} - e^{-\alpha\delta} \zeta_1^T \mathcal{R} \zeta_1 - e^{-\alpha\eta} \zeta_2^T \mathcal{R}_2 \zeta_2, \\ \Phi_{4i} &= \sigma (e_6 - e_8)^T \Omega (e_6 - e_8) - e_8^T \Omega e_8 + \mathbb{H} \left\{ (e_{11} + e_1)^T K_i (e_6 - e_8) \right\} \\ &\quad + \mathbb{H} \left\{ (e_{11} + e_1)^T N_i (-e_{11} - B_i e_1 + A_i e_2 + D_i e_4) \right\}, \\ \Phi_{5i} &= -\mathbb{H} \left\{ (e_2 - \Lambda_1 e_1)^T M_1 (e_2 - \Lambda_2 e_1) + (e_4 - \Lambda_1 e_3)^T M_2 (e_4 - \Lambda_2 e_3) \right\}, \\ \zeta_{11} &= \text{col}\{e_1 - e_9, e_{10} - e_5\}, \zeta_{12} = \text{col}\{e_9 - e_3, e_3 - e_{10}\}, \\ \zeta_1 &= \text{col}\{\zeta_{11}, \zeta_{12}\}, \zeta_2 = \text{col}\{e_1 - e_6, e_6 - e_7\}, \Pi(P_j) = \sum_{j=1}^N \pi_{ij} P_j. \end{aligned}$$

In addition, the control gain is designed by $C_i = N_i^{-1} K_i, i \in \mathfrak{N}$.

Proof The detail of proof is omitted here since the page limits.

4 Numerical Example

In this section, we aim to demonstrate the feasibility and validity of the obtained result in this paper by a numerical example.

Example 1 Consider the systems (7) with the following parameters [12]

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, A_1 = \begin{bmatrix} 2 & -0.1 \\ -5 & 3 \end{bmatrix}, D_1 = \begin{bmatrix} -1.5 & -0.1 \\ -0.2 & -2.5 \end{bmatrix}, \tag{19}$$

$$B_2 = \begin{bmatrix} 0.8 & 0 \\ 0 & 1 \end{bmatrix}, A_2 = \begin{bmatrix} 2 & -0.11 \\ -5 & 3.2 \end{bmatrix}, D_2 = \begin{bmatrix} -1.6 & -0.1 \\ -0.18 & -2.4 \end{bmatrix}, \tag{20}$$

$$\text{TR: } (\pi_{ij}) = \begin{bmatrix} -3 & 3 \\ 5 & -5 \end{bmatrix}, \Lambda_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \Lambda_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \tag{21}$$

$$\delta_1 = 0.1, \delta_2 = 0.2, \mu_1 = 0.2, \mu_2 = 0.1, \tag{22}$$

$$\alpha = 0.16, \sigma = 0.02, \eta = 0.1, \beta = 0.2. \tag{23}$$

According to Theorem 1 and using the MATLAB LMI toolbox, the control gains matrices can be derived: $\Omega = 1.0e + 05 * \begin{bmatrix} 2.3139 & 0.4016 \\ 0.4016 & 0.7217 \end{bmatrix}$ and

$$C_1 = \begin{bmatrix} -0.2865 & -0.0534 \\ -2.2135 & -3.7376 \end{bmatrix}, C_2 = \begin{bmatrix} -0.3173 & -0.0382 \\ -2.1119 & -3.7730 \end{bmatrix}.$$

Fig. 1 Curve of $x(t)$ without control

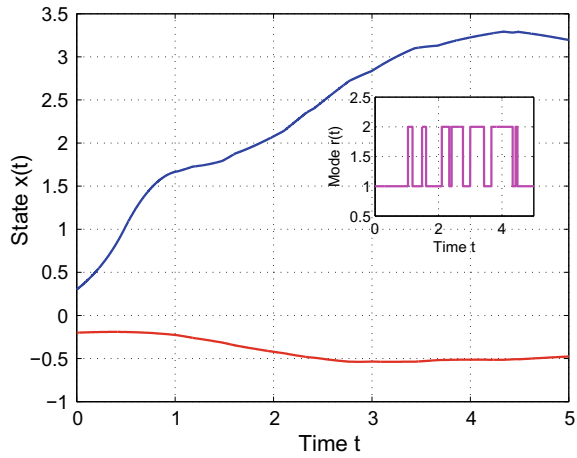
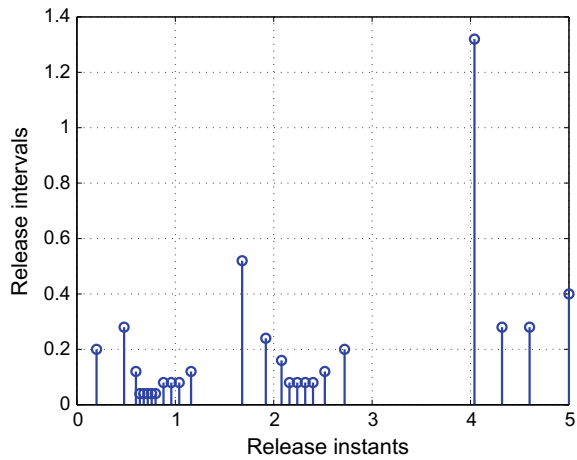


Fig. 2 Release instants and intervals



Under the above parameters, we take the impulsive intensity $q_k = -0.3$, the neuron activation function $f_i(x) = 0.5(|x + 1| - |x - 1|)$, the ATDs $\delta_1(t) = 0.1 + 0.2\sin(t)$, $\delta_2(t) = 0.2 + 0.1\cos(t)$ and the initial value $x(0) = \text{col}\{-0.2, 0.3\}$. Then the feasibility and validity of the obtained result are demonstrated by the following Figs. 1, 2, 3 and 4, and it can be said that more networks resources are saved by using the SEIC scheme. Moreover, comparing the periodic impulsive control scheme, that is, the period of impulsive control is $\eta = 0.1$, then the control numbers will be as high as up to 50, while the control number under the SEIC scheme is 25. Clearly, the control frequency is slashed effectively.

Fig. 3 Curve of $x(t)$ with SEIC

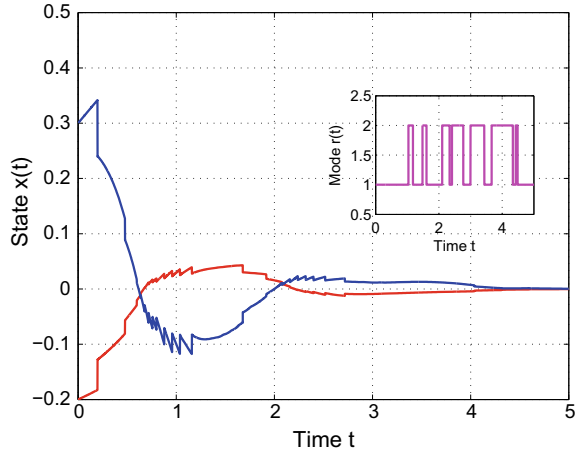
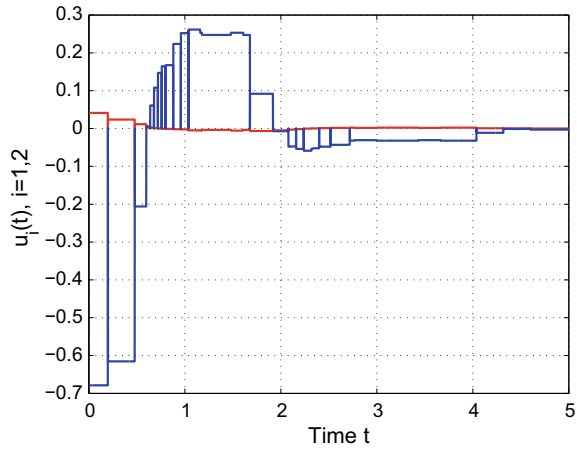


Fig. 4 Control input $u(t)$



5 Conclusion

In this paper, we study the stochastic exponential stabilization problem for MJNNs with ATDs. A novel LKF involving more information about sampled data and ATDs is constructed, and a stochastic exponential stabilization criterion for MJNNs with ATDs is established by employing the SEIC scheme. The feasibility and validity of the obtained result is illustrated by a numerical example, and it is concluded that more networks resources can be saved by using the SEIC scheme.

Acknowledgements This work was funded by the National Natural Science Foundation of China under Grant nos.11671206, 11601474 and 61472093, the China Scholarship Council (CSC), and NSERC Canada.

References

1. Galicki, M., Witte, H., Dörschel, J., Eiselt, M., Griessbach, G.: Common optimization of adaptive preprocessing units and a neural network during the learning period. Application in EEG pattern recognition. *Neural Netw.* **10**(6), 1153–1163 (1997)
2. Ramasamy, S., Nagamani, G., Zhu, Q.: Robust dissipativity and passivity analysis for discrete-time stochastic t-s fuzzy cohen-grossberg markovian jump neural networks with mixed time delays. *Nonlinear Dyn.* **85**(4), 2777–2799 (2016)
3. Zhang, Y., Shi, P., Agarwal, R.K., Shi, Y.: Dissipativity analysis for discrete time-delay fuzzy neural networks with markovian jumps. *IEEE Trans. Fuzzy Syst.* **24**(2), 432–443 (2016)
4. Rawat, A., Yadav, R., Shrivastava, S.: Neural network applications in smart antenna arrays: a review. *AEU-Int. J. Electron. Commun.* **66**(11), 903–912 (2012)
5. Zhang, Y., Shi, Y., Shi, P.: Robust and non-fragile finite-time h-infinity control for uncertain markovian jump nonlinear systems. *Appl. Math. Comput.* **279**, 125–138 (2016)
6. Zhang, Y., Shi, Y., Shi, P.: Resilient and robust finite-time h-infinity control for uncertain discrete-time jump nonlinear systems. *Appl. Math. Model.* **49**, 612–629 (2017)
7. Zhao, Z., Song, Q., He, S.: Passivity analysis of stochastic neural networks with time-varying delays and leakage delay. *Neurocomputing* **125**, 22–27 (2014)
8. Zhao, Y., Gao, H., Mou, S.: Asymptotic stability analysis of neural networks with successive time delay components. *Neurocomputing* **71**(13), 2848–2856 (2008)
9. Cao, J.: Global stability analysis in delayed cellular neural networks. *Phys. Rev. E* **45**(10), 1707–1720 (1999)
10. Tao, L., Qi, L., Sun, C., Zhang, B.: Exponential stability of recurrent neural networks with time-varying discrete and distributed delays. *Nonlinear Anal. Real World Appl.* **10**(4), 2581–2589 (2009)
11. Wang, Z., Liu, Y., Yu, L., Liu, X.: Exponential stability of delayed recurrent neural networks with markovian jumping parameters. *Phys. Lett. A* **356**(4–5), 346–352 (2006)
12. Wu, Z.-G., Shi, P., Su, H., Chu, J.: Stochastic synchronization of markovian jump neural networks with time-varying delay using sampled data. *IEEE Trans. Cybern.* **43**(6), 1796–1806 (2013)
13. Zhu, Q., Cao, J.: Robust exponential stability of markovian jump impulsive stochastic cohen-grossberg neural networks with mixed time delays. *IEEE Trans. Neural Netw.* **21**(8), 1314–1325 (2010)
14. Bao, H., Cao, J.: Stochastic global exponential stability for neutral-type impulsive neural networks with mixed time-delays and markovian jumping parameters. *Commun. Nonlinear Sci. Numer. Simul.* **16**(9), 3786–3791 (2011)
15. Yan, G., Zhou, W., Ji, C., Tong, D., Fang, J.: Globally exponential stability of stochastic neutral-type delayed neural networks with impulsive perturbations and markovian switching. *Nonlinear Dyn.* **70**(3), 2107–2116 (2012)
16. Li, M., Deng, F.: Almost sure stability with general decay rate of neutral stochastic delayed hybrid systems with lvy noise. *Nonlinear Anal. Hybrid Syst.* **24**, 171–185 (2017)
17. Li, Y., Sun, H., Zong, G., Hou, L.: Composite anti-disturbance resilient control for markovian jump nonlinear systems with partly unknown transition probabilities and multiple disturbances. *Int. J. Robust Nonlinear Control* **27**(14) (2016)
18. Li, M., Deng, F.: Necessary and sufficient conditions for consensus of continuous-time multi-agent systems with markovian switching topologies and communication noises. *IEEE Trans. Cybern.* **2**, 1–7 (2019)
19. Peng, C., Li, F.: A survey on recent advances in event-triggered communication and control. *Inf. Sci.* **457**, 113–125 (2018)
20. Åström, K.J., Bernhardsson, B.: Comparison of periodic and event based sampling for first-order stochastic systems. *IFAC Proc. Vol.* **32**(2), 5006–5011 (1999)
21. Tan, X., Cao, J., Li, X.: Consensus of leader-following multiagent systems: a distributed event-triggered impulsive control strategy. *IEEE Trans. Cybern.* **99**, 1–10 (2018)

22. Zhu, W., Wang, D., Liu, L., Feng, G.: Event-based impulsive control of continuous-time dynamic systems and its application to synchronization of memristive neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(8), 3599–3609 (2018)
23. Li, S., Deng, F., Xing, M.: Aperiodic sampled-data robust h-infinity control for delayed stochastic fuzzy systems with quasi-periodical multi-rate approach. *J. Franklin Inst. Eng. Appl. Math.* **356**(8), 4530–4553 (2019)
24. Li, S., Deng, F., Zhao, X.: A new perspective on fuzzy control of the stochastic t-s fuzzy systems with sampled-data. *Sci. China Inf. Sci.* **62**(10) (2019)
25. Zong, G., Ren, H.: Guaranteed cost finite-time control for semi-markov jump systems with event-triggered scheme and quantization input. *Int. J. Robust Nonlinear Control* **29**(15), 5251–5273 (2019)
26. Zhu, Q., Cao, J.: Stability analysis of markovian jump stochastic bam neural networks with impulse control and mixed time delays. *IEEE Trans. Neural Netw. Learn. Syst.* **23**(3), 467–479 (2012)
27. Park, P., Lee, W., Lee, S.: Auxiliary function-based integral inequalities for quadratic functions and their applications to time-delay systems. *J. Franklin Inst.* **352**, 1378–1396 (2015)
28. Park, P., Ko, J., Jeong, C.: Reciprocally convex approach to stability of systems with time-varying delays. *Automatica* **47**(1), 23–238 (2011)

Computational Methods for Differential Equations

Development of a Lattice Boltzmann Model for the Solution of Partial Differential Equations, A Performance Comparison Study with that of the Finite Difference Method



Mahmud Ashrafizaadeh and A. Ghavaminia

Abstract The lattice Boltzmann method (LBM) has attracted much attention in recent years as a recent efficient solution method for fluid flow simulations as well as general PDEs. Due to the local nature of the computations in the lattice Boltzmann method and its ease of programming, the LBM is an ideal candidate for developing efficient parallel PDE solvers suitable for recent computer hardware. In the present study, we have used the lattice Boltzmann method for solving the transient heat diffusion equation. The performance of this method is compared with that of the traditional finite difference based PDE solver. All these solvers have been developed using the Julia programming language, which is a recent player amongst the scientific computing languages. Several benchmark problems in the field of transient heat transfer described by parabolic PDEs are solved, and the results obtained from the aforementioned methods are compared with each other. It is shown that by using the lattice Boltzmann method, it is possible to solve these partial differential equations more efficiently while maintaining the accuracy of the solution.

Keywords Lattice Boltzmann method · Finite difference · Partial differential equation · Numerical performance comparison

1 Introduction

The lattice Boltzmann method (LBM) is a rather young and promising method for simulating complex fluid flow physics. In comparison with the conventional methods in computational fluid dynamics (CFD), LBM is easy for programming, intrinsically

M. Ashrafizaadeh (✉) · A. Ghavaminia
Department of Mechanical Engineering, Isfahan University of Technology,
Isfahan 841568311, Iran
e-mail: mahmud@cc.iut.ac.ir

A. Ghavaminia
e-mail: a.ghavami@alumni.iut.ac.ir

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_24

255

parallel [1, 19], and it is easy to incorporate complicated boundary conditions such as those encountered in porous media problems [6, 10].

In the past years, in addition to solving the fluid flow problems, more applications have been introduced for the lattice Boltzmann method. It has been shown that by modifying a typical LBM, it is possible to solve partial differential equations as well [3, 9, 14, 18, 20]. Solving partial differential equations is required in a vast verity of applications varying from image denoising using the Laplace equation [15] to simulating the heat transfer phenomena in solid or fluid media.

With advances in the high-performance supercomputers and the invention of new GPU acceleration methods, the LBM has gained even more attention due to its intrinsic parallelism and locality of calculations, which is an advantage of the LBM over the conventional partial differential equation solvers [1, 19].

In this study, the accuracy and computational performance of a LBM PDE solver have been investigated and compared with those of a traditional FD PDE solver by solving several benchmarks in the field of transient heat transfer which are described by parabolic PDEs.

2 Description of the Numerical Methods

The transient heat diffusion equation can be written as:

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \quad (1)$$

In which T represents temperature, α is the heat diffusion coefficient, t is time, and x is the spatial direction of the diffusion [8]. One of the well known traditional FD solutions for this PDE is to use the Forward Euler method and the second-order central difference scheme to discretize the Eq. (1) as it is presented in Eq. (2).

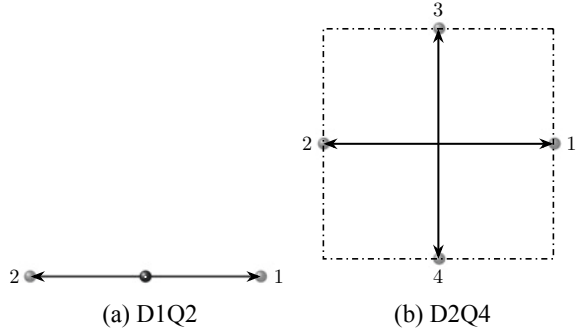
$$T_i^{n+1} = T_i^n (1 - \omega) + \omega(0.5T_{i+1}^n + 0.5T_{i-1}^n) \quad (2)$$

where $\omega = (2\alpha \Delta t)/(\Delta x^2)$. It is necessary to mention that this discrete formulation has a stability criterion ($\Delta t \leq \Delta x^2/2\alpha$), which limits the maximum possible time step size for a fixed Δx . The stability criterion may also vary according to the employed boundary condition type [8]. This method only requires information from the nearest neighboring sites at each time step. Therefore, it is very similar to the LBM in this regard.

The lattice Boltzmann general formulation is:

$$\frac{\partial f}{\partial t} + c \cdot \nabla f = \Omega \quad (3)$$

Fig. 1 Lattice configuration and naming convention used for discretization of the LBM



where f is the population distribution function (PDF), c is the lattice speed, ∇ is the gradient operator, and the Ω is the collision operator [13]. The Ω , in general, is a sophisticated integral, which is hard to compute. There have been several approximations for this integral which up to now, the most widely used and known of them is the Bhatnagar-Gross-Krook (BGK) model [11]. Among the different LBM methods, the BGK collision method [16] is mathematically the simplest. However, it has some deficiencies; for example, the BGK LBM suffers from numerical instability, especially for the simulation of low viscosity fluids [12]. To overcome these deficiencies, several advanced methods have been proposed, among which the multi-relaxation LBM [4] and cascaded central moments method [5] are two of the most known ones.

For discretizing the LBM formulation, Eq. (1), over a domain, there is a need to choose a lattice configuration. In this work, we have chosen the D1Q2 and D2Q4 lattice configuration for one-dimensional and two-dimensional simulations as they are shown in Fig. 1a, b, respectively.

By discretizing the LBM formulation, we will have:

$$\frac{f_k(x, t + \Delta t) - f_k(x, t)}{\Delta t} + c_k \cdot \frac{f_k(x + \Delta x, t + \Delta t) - f_k(x, t + \Delta t)}{\Delta x} = \Omega_k \quad (4)$$

The BGK approximation is:

$$\Omega_k = \frac{1}{\tau} [f_k(x, t) - f_k^{eq}(x, t)] \quad (5)$$

By substitution of the BGK approximation, Eq. (5), into Eq. (4) the BGK LBM becomes:

$$f_k(x + \Delta x, t + \Delta t) - f_k(x, t) = -\frac{\Delta t}{\tau} [f_k(x, t) - f_k^{eq}(x, t)] \quad (6)$$

where f^{eq} is the equilibrium distribution function and τ is the relaxation time factor. By performing the Chapman-Enskog expansion [17] the relation between macroscopic values and the f^{eq} and τ can be determined as:

$$f_k^{eq} = w_i T(x, t) \quad (7)$$

where w_i is the lattice weight factor. For the D1Q2 lattice configuration $w_i = 0.5$, $i = 1, 2$ and for D2Q4 lattice configuration $w_i = 0.25$, $i = 1, \dots, 4$ [13]. Additionally, the temperature T is calculated using the Eq. (8) [13].

$$T(x, t) = \sum_{i=1}^q f_i(x, t) \quad (8)$$

It is more convenient to separate the local and nonlocal part of Eq. (6) and perform the calculations in the collision, Eq. (9), and the streaming, Eq. (10), processes separately.

$$f_k^{post} = f_k(x, t) - \frac{\Delta t}{\tau} [f_k(x, t) - f_k^{eq}(x, t)] \quad (9)$$

$$f_k(x + \Delta x, t + \Delta t) = f_k^{post}(x, t) \quad (10)$$

Boundary Cconditions:

To implement the Dirichlet boundary condition, we should write the flux balance at the considered boundary. For example, at the left boundary, we have:

$$f_1^{eq}(x, t) - f_1(x, t) + f_2^{eq}(x, t) - f_2(x, t) = 0 \quad (11)$$

In Eq. (11) the only unknown is the f_1 .

In the case of the Neumann boundary condition on the left side:

$$q' = -k \frac{T(1) - T(0)}{dx} \quad (12)$$

Substituting $T(1) = f_1(1) + f_2(1)$ and $T(0) = f_1(0) + f_2(0)$ in Eq. (12) and solving for $f_1(0)$ gives:

$$f_1(0) = f_1(1) + f_2(1) - f_2(0) + \frac{q' dx}{k} \quad (13)$$

Finally, in the case of the Robin boundary condition, we have:

$$-\lambda \frac{\partial T(x, t)}{\partial x} = \beta [T(x, t) - T_a] \quad (14)$$

where λ is the solid medium thermal conductivity, β is the convection coefficient in lattice units, and T_a is the ambient temperature [13]. Expanding the Eq. (14) for the right boundary condition yields to:

$$-\lambda \frac{T_n^{p+1} - T_{n-1}^{p+1}}{\Delta x} = \beta [T_n^{p+1} - T_a] \quad (15)$$

This means:

$$f_{1,n}^{post} + f_{2,n}^{post} = \frac{\lambda}{\lambda + \beta \Delta x} (f_{1,n-1}^{post} + f_{2,n-1}^{post}) + \frac{\beta \Delta x}{\lambda + \beta \Delta x} T_a \quad (16)$$

On the right side boundary, the only unknown distribution function is $f_{2,n}$, which can be calculated using Eq. (16).

For a more detailed derivation of boundary conditions, interested readers may refer to references [7, 9, 13, 20].

3 Results

The results presented here are obtained using codes which have been implemented in the Julia programming language [2]. The simulations are performed on a computer system with the following configuration: CPU model: AMD Opteron(tm) Processor 6174, CPU MHz: 2200, L1 cache: 64K, L2 cache: 512K, L3 cache: 5118K, and 94 GB of RAM.

To compare the performance and accuracy of the methods, we used several common one-dimensional and two-dimensional benchmarks introduced in [7, 9, 13]. In the first case a one-dimensional layer of steel with a thermal conductivity of, λ , 35 [W/mK], thermal diffusivity of, α , 7.1795×10^{-6} [m²/s], and thickness of $L = 0.05$ [m] has been considered. The initial condition, the left boundary condition, and the right boundary condition are specified by $T(x, 0) = 0$ [°C], $T(t, 0) = 0$ [°C], and $T(t, L) = 100$ [°C] respectively. To obtain the results illustrated in Fig. 2, the mesh size and time step have been set to $\Delta x = 0.00125$ [m] (*node numbers* (N) = 40) and $\Delta t = 0.001$ [s] respectively. Figure 3 shows the same geometry as the first case (Fig. 2) with the exception of the Robin boundary condition ($\beta = 10$ [W/m²K], $T_a = 20$ [°C]) at the right hand side of the geometry and $T = 150$ [°C] at the left-hand side of the geometry. Figures 2 and 3 clearly show that the results obtained from the LBM and the FD codes are in excellent agreement with each other.

For the two-dimensional formulation, two geometries are selected to compare the performance and the accuracy of the lattice Boltzmann method versus that of the finite difference method. Fig. 4a, b show the geometries, boundary conditions, and initial conditions of the chosen benchmarks.

Figure 5 shows the temperature distribution along the centerline of the case 3 geometry for different dimensionless times. Fig. 6, which is a contour display of the temperature distribution of case 4, shows that the results from both the finite difference method and lattice Boltzmann method are in an excellent agreement with each other. To demonstrate this better, Fig. 7 shows the temperature distributions along the X and Y oriented centerlines of the case 4 geometry. As it is shown in Fig.

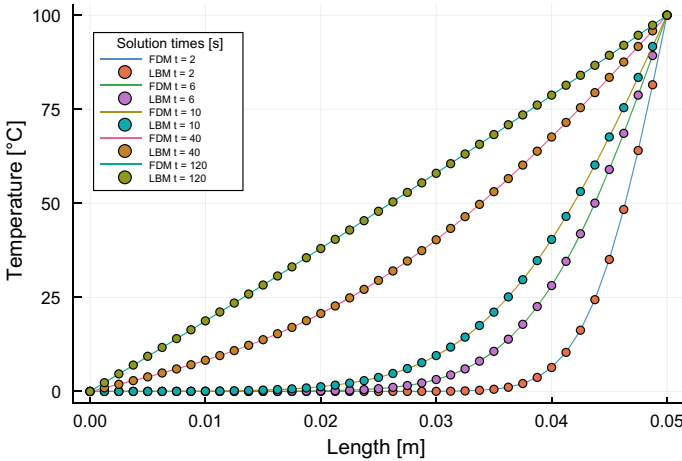


Fig. 2 Temperature distribution at different times for the one-dimensional transient heat transfer with Dirichlet boundary conditions at the left and right sides

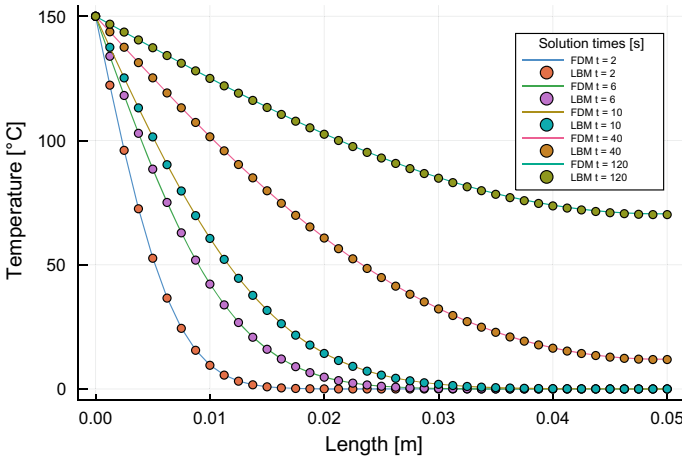


Fig. 3 Temperature distribution at different times for the one-dimensional transient heat transfer with a Dirichlet boundary condition at the left side and a Robin boundary condition at the right side

6 to Fig. 7, the results from the lattice Boltzmann method and the finite difference method are in excellent agreement with each other.

The primary motivation of using the LBM over conventional methods such as the FD for solving differential equations (in this case, the heat diffusion problem) is to benefit from the LBM’s capabilities, particularly the computational performance of the LBM. Another test has been conducted to study the performance of the LBM and the FD, which compares wall clock times required by each method to achieve

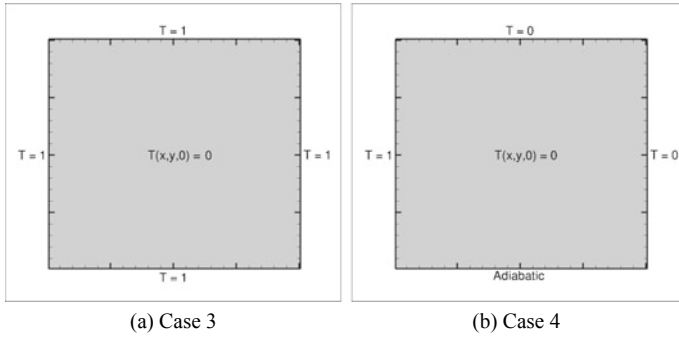


Fig. 4 Boundary conditions and initial condition values for the two-dimensional transient heat transfer case 3 and case 4

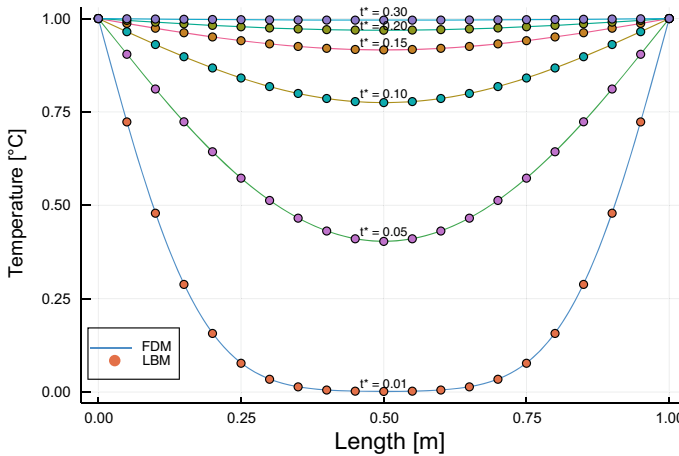


Fig. 5 Temperature distribution along the centerline of the case 3 geometry (Fig. 4a) at different dimensionless times (t^*). Circles and lines represent the LBM and the FD results respectively

a particular result. For this test, we used case 3 geometry to benchmark the results. Times are measured with the benchmarking tool provided by the Julia language (*BenchmarkTools.jl*), which is a tool that can repeat each benchmark and processes the result to eliminate system load noises and to provide a consistent, reliable answer. To conduct a fair comparison, the FD time step has been pushed to its maximum possible value concerning the FD stability limits ($Fo = 0.25$). Then the LBM relaxation time factor has been changed to measure the changes in the performance (Fig. 8).

As it is shown in Fig. 9, the resolution study reveals that even in lower mesh resolutions, the solution is mesh-independent. Nevertheless, we have chosen a finer mesh because we wanted to ensure the accuracy of the study through different simulations,

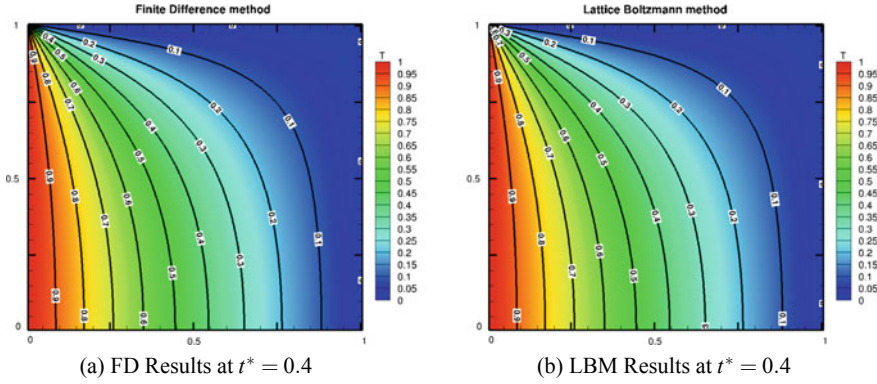


Fig. 6 Case 4 results: Temperature distribution over the solution domain at a dimensionless time ($t^* = 0.4$)

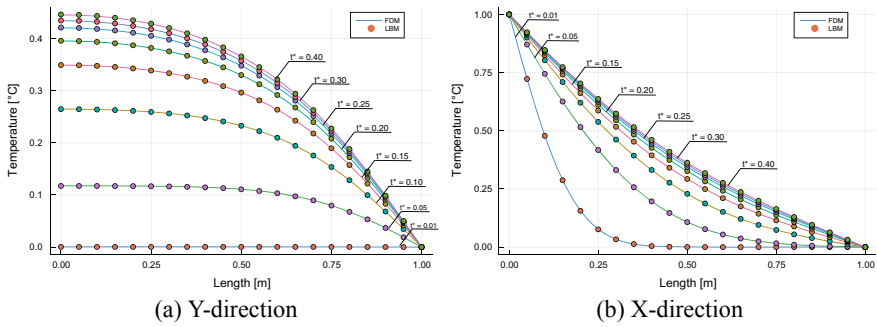


Fig. 7 Temperature distribution along the X and Y oriented centerlines at different dimensionless times

and more importantly, we needed a more time-consuming calculation so that we can provide a more accurate time comparison.

Figure 8 depicts the time ratio of the FD over the LBM calculations, which clearly shows the LBM's superiority over the FD regarding their performances. It is worth mentioning that in this study, we have just implemented a serial code for both the LBM and the FD methods. Due to the parallel nature of the LBM, it is expected that the parallel implementation of these codes would result in even more performance gain in favor of the LBM approach. It is known that by increasing the relaxation time factor of the BGK LBM, its numerical accuracy may deteriorate [12]. To measure this error, the difference between the LBM solution and that of the FD solution for a variety of dimensionless times and LBM relaxation factors are calculated. The results are presented in Fig. 10. The maximum measured difference in our studies has been less than 0.6%, which could be neglected in most common applications. On the other hand, the performance gain is very high. Nearly 0.1 of the computational time is required for the LBM to achieve the same results as that of the FD method. Also, the difference between the LBM and the FD results vanishes by advancing

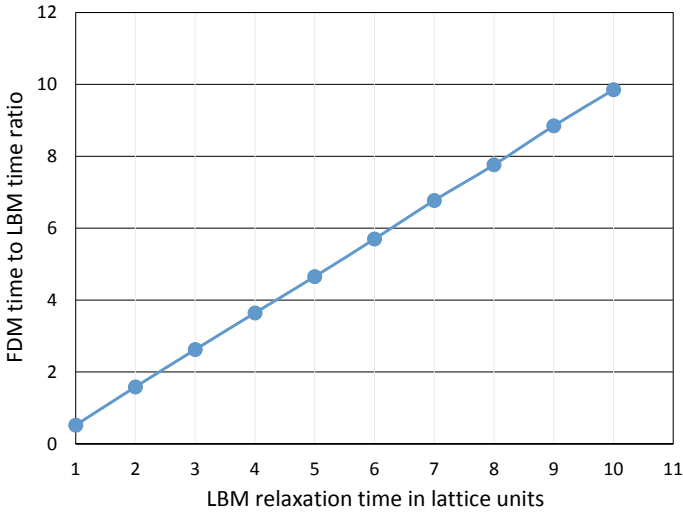


Fig. 8 The time ratio of the FD over the LBM solution. $r = \frac{Time_{FD}}{Time_{LBM}}$

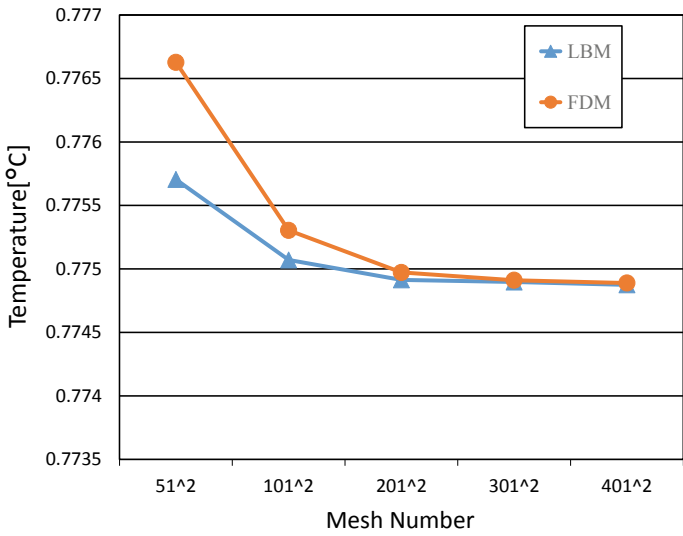


Fig. 9 Mesh study with case 3, LBM relaxation fact 2 and dimensionless time 0.1. The graph shows changes in the temperature values in the center of the geometry for LBM and FD method

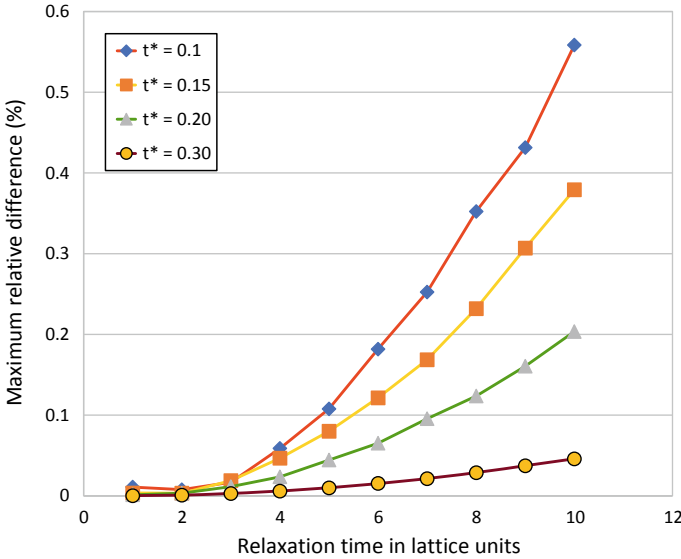


Fig. 10 Maximum relative difference between the FD and LBM solutions through the centerline of the geometry for different LBM relaxation times at different dimensionless times $difference = \max(\frac{T(x,y,t)_{LBM} - T(x,y,t)_{FDM}}{T(x,y,t)_{FDM}}) \times 100$

through time. So it is possible to even achieve the same steady-state results while spending significantly less computational time using the LBM.

4 Conclusion

In this study, the lattice Boltzmann method has been employed to solve the transient heat diffusion equation with different boundary conditions in one and two dimensions. To study the accuracy and performance of the lattice Boltzmann method, several benchmarks have been implemented in the Julia programming language. The results show that the lattice Boltzmann method solution for the transient heat diffusion problems would be as accurate as those obtained by the finite difference method. However, using the lattice Boltzmann method, it is possible to achieve better computational performance and significantly reduce the computational cost. In our benchmarks, the required solution time for the lattice Boltzmann method is an order of magnitude less than that required by the finite difference method for the same level of accuracy.

Acknowledgements The authors gratefully acknowledge the Sheikh Bahaei National High Performance Computing Center (SBNHPCC) for providing computing facilities and time. SBNHPCC

is supported by scientific and technological department of presidential office and Isfahan University of Technology (IUT).

References

1. Ashrafizaadeh, M., Zadehghol, A., Safi, A.: GPU implementation of a lattice boltzmann flow solver. In: 18th Annual International Conference on Mechanical Engineering-ISME2010, number May, Tehran (2010)
2. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**(1), 65–98 (2017)
3. Chai, Z., He, N., Guo, Z., Shi, B.: Lattice Boltzmann model for high-order nonlinear partial differential equations. *Phys. Rev. E* **97**(1), 013304 (2018)
4. D’Humières, D.: Multiple-relaxation-time lattice Boltzmann models in three dimensions. *Philos. Trans. R. Soc. Lond. Ser. A Mathe. Phys. Eng. Sci.* **360**(1792), 437–451 (2002)
5. Geier, M., Greiner, A., Korvink, J.G.: Cascaded digital lattice Boltzmann automata for high Reynolds number flow. *Phys. Rev. E* **73**(6), 066705 (2006)
6. Gharibi, F., Jafari, S., Rahnama, M., Khalili, B., Jahanshahi Javaran, E.: Simulation of flow in granular porous media using combined Lattice Boltzmann method and smoothed profile method. *Comput. Fluids* **177**, 1–11 (2018)
7. Guo, Z., Shu, C.: Lattice Boltzmann method and its applications in engineering. In: *Advances in Computational Fluid Dynamics*, vol. 3. World Scientific (2013)
8. Incropera, F.P., DeWitt, D.P., Bergman, T.L., Lavine, A.S.: *Fundamentals of Heat and Mass Transfer*.pdf, 7 edn. Wiley (2002)
9. Kałuza, G.: The numerical solution of richards equation using the Lattice Boltzmann method. *Appl. Mech. Mater.* **11**(1), 23–30 (2012)
10. Khalili, B., Rahnama, M., Jafari, S., Gharibi, F., Jahanshahi Javaran, E.: Lattice Boltzmann simulation of solid particles motion in a three dimensional flow using smoothed profile method. *J. Appl. Fluid Mech.* **10**(4), 1091–1103 (2017)
11. Krüger, T., Kusumaatmaja, H., Kuzmin, A., Shardt, O., Silva, G., Viggien, E.M.: *The Lattice Boltzmann Method*. Graduate Texts in Physics. Springer International Publishing, Cham (2017)
12. Luo, L.-S., Liao, W., Chen, X., Peng, Y., Zhang, W.: Numerics of the lattice Boltzmann method: effects of collision models on the lattice Boltzmann simulations. *Phys. Rev. E* **83**(5), 056710 (2011)
13. Mohamad, A.A.: *Lattice Boltzmann Method*. Springer, London (2011)
14. Otomo, H., Boghosian, B.M., Dubois, F.: Two complementary lattice-Boltzmann-based analyses for nonlinear systems. *Physica A Stat. Mechan. Appl.* **486**, 1000–1011 (2017)
15. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: *ACM SIGGRAPH 2003 Papers on—SIGGRAPH ’03*, p. 313. ACM Press, New York, NY, USA (2003)
16. Qian, Y.H., D’Humières, D., Lallemand, P., Search, H., Journals, C., Contact, A., Iopscience, M., Address, I.P.: Lattice BGK models for Navier-Stokes equation. *Europhys. Lett. (EPL)* **17**(6), 479–484 (1992)
17. Rosenboum, E.J.: The mathematical theory of non-uniform gases (Chapman, S.; Cowling, T. G.). *J. Chem. Educ.* **18**(1), 48 (1941)
18. Zergani, S., Aziz, Z.A., Viswanathan, K.K.: Exact solutions and lattice Boltzmann method modelling for shallow water equations. *Glob. J. Pure Appl. Math.* **12**(3), 2243–2266 (2016)
19. Zhao, Y.: Lattice Boltzmann based PDE solver on the GPU. *Vis. Comput.* **24**(5), 323–333 (2008)
20. Zhou, Z., Ma, J.: Lattice Boltzmann methods for solving partial differential equations of exotic option pricing. *Front. Math. China* **11**(1), 237–254 (2016)

Using Shooting Approaches to Generate Initial Guesses for ODE Parameter Estimation



Jonathan Calver, Jienan Yao, and Wayne Enright

Abstract We consider the parameter estimation problem for parameterized systems of ordinary differential equations (ODEs). This problem involves finding the set of parameters that best fit a set of observed data. In particular, we consider techniques for generating initial guesses that are sufficiently close to the best fit parameters, so that a shooting approach is likely to converge. We discuss approaches used in the literature and demonstrate how they can be improved using ideas motivated by progressive and multiple shooting. Our proposed approach is then applied to a test problem from the literature.

Keywords Ordinary differential equations · Inverse problems · Shooting methods

1 Introduction

Parameterized initial value problems (IVPs) are used in a wide range of applications, including investigations of population dynamics [3], enzyme kinetics [11], the spread of disease [14], chemical reactions [16], and neuron signalling [10]. An IVP consists of a system of ordinary differential equations (ODEs) and a set of initial conditions, which specify the initial state of the model. The solution of an IVP can be approximated by simulating the model state from the initial time to some final time of interest. The solution of the IVP over an interval of interest is referred to as a trajectory. Parameter estimation seeks to find the set of model parameters such that the model best fits the observed data, as defined by an appropriate objective function.

J. Calver (✉) · J. Yao · W. Enright
Department of Computer Science, University of Toronto, 40 St. George Street,
Toronto, ON M5S 2E4, Canada
e-mail: calver@cs.toronto.edu

J. Yao
e-mail: jnyao@cs.toronto.edu

W. Enright
e-mail: enright@cs.toronto.edu

Estimating the best fit parameters can be a computationally intensive task. Evaluating the objective function for a candidate set of model parameters requires a trajectory simulation. If the model is complex, or the interval of interest is large, this simulation can be quite time consuming. This observation has led to a variety of techniques being proposed to reduce the number of model trajectory simulations required to approximate the best fit model parameters. In this work, we describe some of these techniques, propose a modification to one of the techniques based on the ideas of progressive and multiple shooting, and demonstrate its use in a two stage parameter estimation procedure.

1.1 Definitions and Notation

We consider the parameterized initial value problem (IVP),

$$y'(t) = f(t, y(t), p), \quad y(0) = y_0, \quad t \in (0, T), \quad (1)$$

where $y(t)$ is the state vector of dimension n_y , p is a constant vector of model parameters of dimension n_p , and y_0 are the initial conditions of the state vector, $y(0)$. We will denote the solution of (1) by $y(t, p)$. In some applications, the parameters only appear linearly in f and one can often exploit this structure to rewrite f as,

$$f(t, y(t), p) = G(t, y(t))p. \quad (2)$$

In the rest of this paper, we consider this case, where all of the parameters appear linearly.

1.2 Least Squares Parameter Estimation

We assume that a set of observations of the entire state vector is known and given by,

$$\hat{y}_j(t_i) = y_j(t_i) + \mathcal{N}(0, \sigma_{ij}^2), \quad \text{for } i = 1, \dots, n_o \quad j = 1, \dots, n_y, \quad (3)$$

where n_o is the number of observation points, $y_j(t_i)$ denotes the j th component of the true state vector at time t_i , and $\mathcal{N}(0, \sigma_{ij}^2)$ is normally distributed noise with variance σ_{ij}^2 . Given such data, parameter estimation is often performed using maximum likelihood estimation (MLE). This leads to a nonlinear least squares objective function,

$$L(p) = \sum_{i=1}^{n_o} \sum_{j=1}^{n_y} \frac{(\hat{y}_j(t_i) - y_j(t_i, p))^2}{2\sigma_{ij}^2}. \quad (4)$$

The parameter estimation problem we consider is to find,

$$\hat{p} = \arg \min_p L(p), \quad \text{subject to (1)}. \quad (5)$$

1.3 Shooting Approaches

In this work, we consider the single shooting (SS) approach (see for example [1]) for solving this problem. In the single shooting approach, an IVP solver is used to approximate a trajectory, $y(t, p)$, to within a user specified tolerance, whenever we evaluate the objective function. That is, we approximate the solution of (1) on each iteration of the optimization.

In order to find \hat{p} using a gradient based optimizer and the single shooting approach, we require approximations to the sensitivity information. This requires us to approximate the model sensitivities, $\frac{dy}{dp}(t)$, at each observation point, t_i . For example, the model sensitivities can be approximated by simulating the variational equations simultaneously with the original system of ODEs.

A common criticism of using a gradient based optimizer and the single shooting approach is that it relies on an initial p , call it p_o , that is sufficiently close to the best fit \hat{p} . If p_o is not close enough to \hat{p} , the optimization may converge to a poor, local minimum, converge slowly, or fail to converge at all. Two modifications to single shooting have been proposed that make it less sensitive to the choice of p_o .

Incremental shooting [19] and progressive shooting (PS) [15] are both based on the observation that if the parameters are too far from their true values, then the solution of the IVP may only remain close to the data near the initial time. Trying to fit to data further in the simulation can lead to convergence to a poor local minimum or the simulation might fail before reaching $t = T$ for certain values of the parameters. Progressive shooting proceeds by first fitting to the data over a shorter interval, say $[0, \tilde{T}]$, then progressively increasing the length of the interval. This can also be viewed as a form of continuation, where the continuation parameter is \tilde{T} .

Multiple shooting (MS) is a more robust and better conditioned form of single shooting, which is widely used for the numerical solution of ODE boundary value problems (BVPs) (see, for example [4, 15]). This technique is particularly effective if the problem is not well conditioned. It has also been suggested for estimating the parameters in systems of ODEs [9, 20].

In multiple shooting, the interval over which the system is simulated is divided into N_{MS} subintervals. Additional parameters are added to specify the state vector at the beginning of each subinterval. Equality constraints are introduced at the boundary of each subinterval and these additional constraints are added to the objective function, (4).

A significant advantage of this approach is that it allows discontinuities in the intermediate trajectories (i.e. violation of the introduced equality constraints) to exist during the optimization. Also, since it restarts the simulations at the beginning of each subinterval, it is less likely that the IVP solver will fail. When used in conjunction

with a gradient based optimizer, a downside of multiple shooting is that each iteration is more computationally expensive than single shooting; although we note that, on each iteration, each of the simulations over the N_{MS} subintervals can be performed in parallel.

Alternatively, we can try to determine a value for p_o , such that SS is likely to converge to \hat{p} , which is the approach we will consider here. The techniques described in [5], hybrid optimizers [21], and ACCEL [7] can all be viewed as two stage approaches, where first a suitable p_o is found, then SS with a gradient based optimizer is used to determine \hat{p} . We now discuss ways to use the observed data and structure of the model to obtain such a suitable p_o .

1.4 Obtaining a suitable p_o

The expensive part of evaluating the objective function, (4), is simulating the underlying ODE IVP. Varah [22] and others [2, 7, 12] recognized that if one uses the observed values of $y(t)$ to approximate $y'(t)$, then one can formulate a related least squares problem,

$$\min_p \int_0^T \|(\tilde{y}'(t) - f(t, \tilde{y}(t), p))\|^2 dt, \quad (6)$$

where $\tilde{y}(t)$ is an approximation of $y(t)$ over the interval, $[0, T]$, based on the observed data. Note, in the work of Varah [22], a sum over the observation times was used, rather than an integral. An estimate for p obtained in this way is referred to as a smooth and match estimator (SME) [12]. In terms of computation, a major benefit of SME is that the numerical derivatives to be approximated can be significantly less expensive to compute than a simulation of an ODE IVP. We also do not have to worry about what happens when a set of parameters would cause the simulation to fail.

More recently, Dattner [8] has suggested a related approach using the integral form of the associated ODE IVP. The resulting approach is referred to as integral SME (INT-SME). We note that this approach is similar to those that use the Collage Theorem [17, 18]. If we assume the structure of the ODE IVP in (2), then the associated least squares problem defining INT-SME is given by,

$$\min_p \int_0^T \left\| \tilde{y}(t) - \left(y_0 + \left[\int_0^t G(\tau, \tilde{y}(\tau)) d\tau \right] p \right) \right\|^2 dt. \quad (7)$$

Note that this least squares problem is linear in p (and y_0) and unlike SME, this does not require $y'(t)$ to be explicitly approximated.

The above methods can be extended to the general case with nonlinear parameters and this is discussed in [5]. It is also sometimes possible to extend these methods to handle unobserved states [5, 6].

2 Proposed method

While INT-SME can be quite effective, it may produce poor results if the available data doesn't allow for the underlying shape of the trajectory to be recovered with enough accuracy. Since the estimator relies on cumulative integrals, errors may propagate over the interval. As a simple example, consider an undamped oscillator,

$$y'(t) = v(t), \tag{8}$$

$$v'(t) = -ky(t). \tag{9}$$

In this example, we see that if the noise in the observations is Gaussian, then the variance in the only component of $\int_0^t G(\tau, \tilde{y}(\tau)) d\tau$ will grow like t , since it is the integral of $-\tilde{y}(\tau)$. Of course, depending on the form of G , it may be hard to predict how the error will propagate for a given model. This suggests it might be helpful to restart the integrals to prevent too much error from propagating and this motivates our proposed method.

Note that usually the hope is that the smoother can sufficiently reduce the error to mitigate this problem, but if the underlying trajectory contains peaks that are not sufficiently sampled, then the peak might get smoothed out—potentially causing more error to propagate.

First, we define INT-SME(s), where s takes the place of T in the upper limit of integration in the outermost integral in (7). Inspired by progressive shooting, this will only consider the subset of observations between $t = 0$ and $t = s$. We then propose a multiple shooting inspired version of INT-SME(s). This approach considers,

$$\min_{p, \{\bar{y}(t_i)\}_{i=1}^m} \sum_{i=1}^m \left[\int_{t_i}^{t_{i+1}} \left\| \tilde{y}(t) - \left(\bar{y}(t_i) + \left[\int_{t_i}^t G(\tau, \tilde{y}(\tau)) d\tau \right] p \right) \right\|^2 dt \right], \tag{10}$$

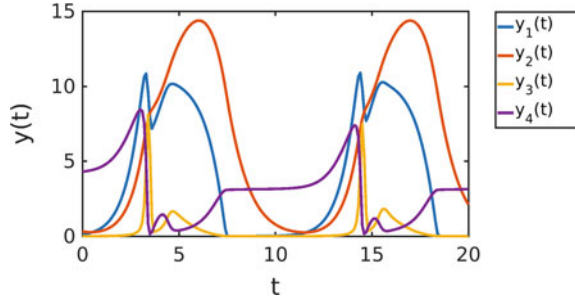
where m is the number of shooting intervals used and the set of t_i 's partition the interval from 0 to s . Note that $m = 1$ corresponds to INT-SME, using the observations up to s . In our numerical experiments, we have used uniform partitions. Unlike true multiple shooting, we do not enforce equality constraints at the end of each shooting interval, since doing so would reduce to INT-SME(s). One could also fix the $\bar{y}(t_i)$'s to their observed values, but in our experiments, we have included the $\bar{y}(t_i)$'s in the linear least squares problem. We will refer to this approach as INT-SME(m, s).

3 Numerical Experiments

We now consider an example from the literature to demonstrate the performance of our proposed method.

The Calcium ion model is a system of ODEs describing the oscillations of Ca^{2+} ions in the cytoplasm of eukaryotic cells, which play a role in cellular information

Fig. 1 True trajectories for the Calcium Ion test problem



processing. For a complete description of this model, see [16]. The model is given by,

$$G^*_\alpha' = k_1 + k_2 G^*_\alpha - k_3 PLC^* \frac{G^*_\alpha}{G^*_\alpha + Km_1} - k_4 Ca_{cyt} \frac{G^*_\alpha}{G^*_\alpha + Km_2}, \quad (11)$$

$$PLC^{*'} = k_5 G^*_\alpha - k_6 \frac{PLC^*}{PLC^* + Km_3}, \quad (12)$$

$$Ca_{cyt}' = k_7 PLC^* Ca_{cyt} \frac{Ca_{er}}{Ca_{er} + Km_4} + k_8 PLC^* + k_9 G^*_\alpha - k_{10} - k_{11} \frac{Ca_{cyt}}{Ca_{cyt} + Km_6}, \quad (13)$$

$$Ca_{er}' = -k_7 PLC^* Ca_{cyt} \frac{Ca_{er}}{Ca_{er} + Km_4} + k_{11} \frac{Ca_{cyt}}{Ca_{cyt} + Km_6}, \quad (14)$$

where the state variables, $y = [G^*_\alpha, PLC^*, Ca_{cyt}, Ca_{er}]$, are concentrations of four compounds, which interact in the calcium-signaling pathway. The parameters are chosen to be $k_1 = 0.09$, $k_2 = 2$, $k_3 = 1.27$, $k_4 = 3.73$, $k_5 = 1.27$, $k_6 = 32.24$, $k_7 = 2$, $k_8 = 0.05$, $k_9 = 13.58$, $k_{10} = 153$, $k_{11} = 4.85$, $Km_1 = 0.19$, $Km_2 = 0.73$, $Km_3 = 29.09$, $Km_4 = 2.67$, $Km_5 = 0.16$, $Km_6 = 0.05$. The nonlinear parameters (the Km 's) are considered fixed and the linear parameters (the k 's) are estimated. The initial conditions are treated as known and are given by $y(0) = [0.12, 0.31, 0.0058, 4.3]$. The model is simulated for $t \in [0, 20]$. For this specific parameterization, the solution exhibits a limit cycle [20]. The true trajectories corresponding to these parameters are shown in Fig. 1.

To generate the noisy data, we simulate the ODE IVP with the true parameter values and take observations every 0.1 time units, from $t = 0$ to $t = 20$. Noise is added relative to the magnitude of each component of the state vector, such that each observation has roughly 6.5% error. This is the same experimental setup used in [20], in which the authors demonstrated that multiple shooting can be more robust than single shooting when a good initial guess is not available. They generated one set of noisy data (as we described) and ran each of simple shooting (SS) and multiple shooting (MS) with $N_{MS} = 17$ from 250 random initial guesses on the model parameters, which were drawn uniformly from $[0, 1]^p$. They found that SS converged to \hat{p} only 4% of the time and MS converged to \hat{p} 49% of the time.

Table 1 The percentage of times the final optimization succeeded when INT-SME(m, s) was used to generate p_o . Note that $m = 1$ corresponds to INT-SME

m/s	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	96	98	98	92	93	85	95	92	94	95	92	84	72	66	65	63
2	93	100	92	90	98	98	98	100	95	95	98	91	80	66	55	64
4	98	100	98	92	97	100	100	99	98	97	98	96	94	93	86	91
8	98	100	99	100	100	100	99	99	100	100	100	99	98	99	98	100
16	99	100	100	98	100	100	100	100	100	100	100	100	100	100	99	100

For our experiments, we generated 200 sets of noisy data as described above and attempted to generate a suitable p_o for each set of data. As noted in [13], the choice of smoother can bias the estimates generated by procedures like SME and INT-SME. This is the case in this example, due to sharp peaks in the true trajectory. For our numerical experiments, we found that reasonable results were obtained by not using a smoother to ensure that the peaks are preserved. To efficiently approximate the cumulative integrals required by INT-SME(m, s), we used the trapezoidal rule—with the same mesh as the observed data.

Given these initial guesses, we then performed the final optimization, (5), using Matlab’s implementation of Levenberg-Marquardt in lsqnonlin. We used the DDEM package [23] as our IVP solver to simulate the model trajectories and their associated variational equations to approximate the required sensitivities. A tolerance of 10^{-5} was used for all simulations. We consider the method to have succeeded if $L(\hat{p})$ is less than the objective function evaluated for the true parameter vector. Since a relative error model is used in this example, we let $\sigma_{ij} = 0.065|\hat{y}_j(t_i)|$. We found that this sometimes led to issues in the early iterations of the optimizations for components of $y(t)$ close to zero, so we used a constant σ_{ij} for the first few iterations.

Table 1 shows the convergence results for our experiments where we varied m and s . We see that increasing m seems to increase the success rate, although there are still occasional failures. Most importantly, we observe that for $m \in [1, 2]$, we obtain significantly worse results if we include observations past around $t = 16$. Recall that $m = 1$ corresponds to INT-SME(s)—with $s = 20$ being the original INT-SME, where all observations are included.

Table 2 shows the timing results. We observe that the time taken in all cases is quite similar, since the cost is dominated by the cost of the final optimization. When a better initial guess is obtained, the final optimization may take fewer iterations—resulting in a reduction in cost. We observe that when s is large, increasing m somewhat reduces the time taken. The time taken decreases with s , but for smaller values of s , increasing m has less impact on the time taken.

Table 2 Average cost (in seconds) of the full procedure using INT-SME(m, s) to generate p_o and using SS to obtain \hat{p} . Note that $m = 1$ corresponds to INT-SME

m/s	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.87	0.84	0.92	0.89	1.00	0.85	0.83	0.98	0.91	0.85	0.86	1.01	1.45	1.48	1.31	1.65
2	0.93	0.84	0.85	0.93	0.91	0.86	0.89	0.93	1.05	0.80	0.93	1.18	1.17	1.55	1.32	1.57
4	0.91	0.77	0.84	0.97	0.80	0.83	0.83	0.89	0.86	0.85	0.93	0.92	1.04	1.08	1.31	1.19
8	0.87	0.81	0.86	0.86	0.81	0.81	0.83	0.81	0.82	0.85	0.85	0.86	0.94	0.98	0.90	0.95
16	0.98	0.92	0.93	0.89	0.90	0.92	0.92	0.95	0.92	0.95	0.91	0.96	0.98	1.04	1.01	1.03

3.1 Comparison to random sampling

We also include an experiment similar to that in [20], in order to see the relative performance of our proposed methods and to compare SS and PS, rather than SS and MS as was done in [20]. We summarize the results in Table 3.

As we can see, random sampling is much less likely to succeed and will also be slower than if we use any of the methods in Table 2. We also observe that PS is almost twice as likely to converge to the global minimum and it is actually faster as well. This can be explained by the fact that PS starts with shorter trajectories and only attempts longer trajectories once the parameters are fitting the initial data values reasonably well. We do note that when we tried PS with a longer initial interval, we observed slightly better convergence, but with a cost more similar to SS than PS with a shorter initial interval.

For a more direct comparison, we can do a simple calculation based on the probability of success and the cost per guess for random sampling. From Table 2, we have that our INT-SME(m, s) procedure takes roughly one second and is very likely to succeed. From this experiment, we have that for a given data set, random sampling succeeds roughly 10% of the time and takes about 4s. So on average, we would expect to require about 40s of computation (or about ten samples). Of course, these samples could be checked in parallel or a more sophisticated sampling method could be used to reduce this cost. Similarly, our PS implementation that took 2.5s per guess and had a success rate of 20%, would require about 12.5s of computation (or about five samples) on average.

Table 3 Timing and success rate for the shooting approaches applied to the Calcium Ion test problem. PS($\vec{T} = 10, 20$) means we first fit to the first half of the data and then used that estimate to fit to all of the data

Method	Success rate	Average time (s)
SS	$\frac{27}{250}$	3.9
PS($\vec{T} = 5, 10, 15, 20$)	$\frac{50}{250}$	2.5
PS($\vec{T} = 10, 20$)	$\frac{65}{250}$	3.7

4 Conclusions

A modified version of INT-SME inspired by multiple and progressive shooting was proposed. Its performance was demonstrated on a test problem from the literature, for which single shooting was known to perform poorly. Our proposed method, INT-SME(m, s), was shown to both improve the robustness and reduce the cost of a two stage, gradient based single shooting ODE IVP parameter estimation procedure. In future work we will further test our proposed method on other problems from the literature.

Acknowledgements We would like to express our gratitude to the anonymous reviewers for having carefully read of our paper and providing us with constructive feedback. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Bard, Y.: Nonlinear Parameter Estimation. Academic Press (1974)
2. Bellman, R., Roth, R.: The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.* **34**(1), 26–33 (1971)
3. Berryman, A.A.: The origins and evolution of predator-prey theory. *Ecology* **73**(5), 1530–1535 (1992)
4. Bock, H., Plitt, K.: A multiple shooting algorithm for direct solution of optimal control problems. In: Proceedings 9th IFAC World Congress Budapest, pp. 243–247 (1984)
5. Calver, J.: Parameter estimation for systems of ordinary differential equations. Ph.D. thesis, University of Toronto (2019)
6. Dattner, I.: A model-based initial guess for estimating parameters in systems of ordinary differential equations. *Biometrics* **71**(4), 1176 (2015)
7. Dattner, I., Gugushvili, S.: Accelerated least squares estimation for systems of ordinary differential equations. [arXiv:1503.07973](https://arxiv.org/abs/1503.07973) (2015)
8. Dattner, I., Klaassen, C.: Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Statist.* **9**(2), 1939–1973 (2015)
9. van Domselaar, B., Hemker, P.: Nonlinear parameter estimation in initial value problems. *Stichting Mathematisch Centrum. Numerieke Wiskunde (NW 18/75)* (1975)
10. FitzHugh, R.: Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1**(6), 445–466 (1961)
11. Goodwin, B.: Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.* **3**, 425–428 (1965)
12. Gugushvili, S., Klaassen, C., et al.: \sqrt{n} -consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli* **18**(3), 1061–1098 (2012)
13. Hooker, G.: Forcing function diagnostics for nonlinear dynamics. *Biometrics* **65**(3), 928–936 (2009)
14. Kermack, W., McKendrick, A.: Contributions to the mathematical theory of epidemics. Part I. *Proc. R. Slat. Soc. A* **115**, 700–721 (1927)
15. Krogh, F., Keener, J., Enright, W.: Reducing the number of variational equations in the implementation of multiple shooting. *Numer. Bound. Value ODEs* 121–135 (1985)
16. Kummer, U., Olsen, L.F., Dixon, C.J., Green, A.K., Bomber-Bauer, E., Baier, G.: Switching from simple to complex oscillations in calcium signaling. *Biophys. J.* **79**(3), 1188–1195 (2000)

17. Kunze, H., Hicken, J., Vrscay, E.: Inverse problems for odes using contraction maps and suboptimality of the collage method. *Inverse Prob.* **20**(3), 977 (2004)
18. Kunze, H., Vrscay, E.: Solving inverse problems for ordinary differential equations using the picard contraction mapping. *Inverse Prob.* **15**(3), 745 (1999)
19. Michalik, C., Hannemann, R., Marquardt, W.: Incremental single shooting—a robust method for the estimation of parameters in dynamical systems. *Comput. Chem. Eng.* **33**(7), 1298–1305 (2009)
20. Peifer, M., Timmer, J.: Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *IET Syst. Biol.* **1**(2), 78–88 (2007)
21. Rodriguez-Fernandez, M., Mendes, P., Banga, J.: A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* **83**(2), 248–265 (2006)
22. Varah, J.: A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.* **3**(1), 28–46 (1982)
23. Zivari-Piran, H.: Efficient simulation, accurate sensitivity analysis and reliable parameter estimation for delay differential equations. Ph.D. thesis, University of Toronto (2009)

A Computational Study for Solving Inverse Problems for Mixed Variational Equations on Perforated Domains



A. I. Garralda-Guillem, Herb Kunze, Davide La Torre, and M. Ruiz Galán

Abstract In this paper we give some conditions for the existence of solution of a system of mixed variational equations and of a related inverse problem. We also conduct a computational study related to the collage-based approach for solving inverse problems for mixed variational equations on perforated domains.

Keywords Mixed variational equations · Inverse problems · Perforated domains

1 Introduction

Mixed variational formulations are very well known and stated area in numerical analysis and treatment of Partial Differential Equations (see, for instance, [3] and some of its generalizations [5, 6]).

In this paper we consider a perturbed variational equation in which a perturbation term, expressed in terms of a bilinear form, is added to the basis model. We focus on its formulation on a perforated domain, that is a domain that shows the presence of holes. This problem can be now analyzed from both a direct and inverse approach: the direct problem is the analysis of the properties of existence, uniqueness, and

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

M. Ruiz Galán
e-mail: mruizg@ugr.es

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

D. La Torre
SKEMA Business School—Universite' de la Cote-d'Azur, Sophia Antipolis Campus,
Valbonne, France
e-mail: davide.latorre@skema.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_26

stability of the solution. This is also referred to the notion of well-posedness in the sense of Hadamard [7].

The inverse problem, instead, aims to identify causes from effects. In practice, this may be done by using observed data to estimate parameters in the functional form of a model. Usually an inverse problem is ill-posed because some of the properties related to existence, uniqueness, and stability fail to hold. When this happens, it is crucial to identify a suitable numerical scheme that ensures the convergence to the solution.

The literature is rich in papers studying ad hoc methods to address ill-posed inverse problems by minimizing a suitable approximation error along with utilizing some regularization techniques [8, 17–21]. Many inverse or parameter identification problems can be viewed in terms of the approximation of a target element in a complete metric space by the fixed point of a contraction mapping and by means of *Banach's Theorem* and the so-called *Collage Theorem* [1]. This approach has been used a lot in fractal imaging to approximate a target image by the fixed point (image) of a contractive fractal transform [1, 14].

These ideas have been extended to inverse problems for ordinary differential equations and their application to different fields in [9, 11]. Fractal-based methods have been also extended to solving inverse problems for partial differential equations over solid and perforated domains. [2, 5, 6, 10, 12, 13, 15, 16].

The paper is organized as follows. Section 2 presents the basics on inverse problems using an extension of the Collage Theorem, known as Generalized Collage Theorem. Section 3 presents a general formulation of a system of mixed variational equations derived in [4] and illustrates a fourth-order differential equations that can be written as a mixed-variational equation. Section 4 illustrates some computational studies, and finally Sect. 5 concludes.

2 Basics on Inverse problems for Variational Equations using the Generalized Collage Theorem

Let E be a Hilbert space, and consider the following variational equation: Find $u \in E$ such that

$$a(u, v) = x^*(v), \quad (1)$$

for any $v \in E$, where $x^*(v)$ and $a(u, v)$ are linear and bilinear maps, respectively, both defined on a Hilbert space E . Let $\langle \cdot \rangle$ denote the inner product in E , $\|u\|^2 = \langle u, u \rangle$ and $d(u, v) = \|u - v\|$, for all $u, v \in E$. The existence and uniqueness of solutions to this kind of equation are provided by the classical Lax-Milgram representation theorem. The following theorem presents how to determine the solution to the inverse problem for the above variational problem. Following our earlier studies of inverse problems using fixed points of contraction mappings, we shall refer to it as a “generalized collage method.”

Theorem 1 (Generalized Collage Theorem) [12] For any $\lambda \in \Lambda$, let $a_\lambda : E \times E \rightarrow \mathbb{R}$ be a family of bilinear forms and $x_\lambda^* : E \rightarrow \mathbb{R}$ be a family of linear forms, and suppose that

1. there exists a constant $M = \sup_{\lambda \in \Lambda} M_\lambda > 0$ such that for any $\lambda \in \Lambda$, $|a_\lambda(u, v)| \leq M_\lambda \|u\| \|v\|$ for all $u, v \in E$, and
2. there exists a constant $m = \inf_{\lambda \in \Lambda} m_\lambda > 0$ such that for any $\lambda \in \Lambda$, $|a_\lambda(u, u)| \geq m_\lambda \|u\|^2$ for all $u \in E$.

Then, according to the Lax-Milgram theorem, for any $\lambda \in \Lambda$ there exists a unique vector u_λ such that

$$a_\lambda(u_\lambda, v) = x_\lambda^*(v)$$

for all $v \in E$. Then, for any $u \in E$,

$$\|u - u_\lambda\| \leq \frac{1}{m_\lambda} F(\lambda), \tag{2}$$

where

$$F(\lambda) = \sup_{v \in E, \|v\|=1} |a_\lambda(u, v) - x^*(v)| = \|a_\lambda(u, \cdot) - x^*\|. \tag{3}$$

In order to ensure that the approximation u_λ is close to a target element $u \in E$ we can, by the Generalized Collage Theorem, try to make the term $F(\lambda)/m_\lambda$ as close to zero as possible. The inverse problem can be reduced to the minimization of the function $F(\lambda)$ on the space Λ , that is,

$$\min_{\lambda \in \Lambda} F(\lambda). \tag{4}$$

3 Mixed Variational Equations

Mixed-Variational Equations are extensions of the classical variational equations presented in the previous section. The perturbed version adopt this form: Let E and F be real Hilbert spaces, $a : E \times E \rightarrow \mathbb{R}$, $b : E \times F \rightarrow \mathbb{R}$ and $c : F \times F \rightarrow \mathbb{R}$ are continuous bilinear forms, $x^* : E \rightarrow \mathbb{R}$ and $y^* : F \rightarrow \mathbb{R}$ are linear forms. The problem under consideration is the following: Find (w, ψ) such that

$$\begin{cases} v \in E \Rightarrow a(w, v) + b(v, \psi) = x^*(v) \\ \phi \in W \Rightarrow b(w, \phi) + c(\psi, \phi) = y^*(\phi) \end{cases} \tag{5}$$

The following result (see [4]) states a sufficient condition that guarantees existence and uniqueness of the solution.

Theorem 2 Assume that E and F are real Hilbert spaces, Λ is nonempty set and that for all $\lambda \in \Lambda$, $a_\lambda : E \times E \rightarrow \mathbb{R}$, $b_\lambda : E \times F \rightarrow \mathbb{R}$ and $c_\lambda : F \times F \rightarrow \mathbb{R}$ be continuous and bilinear forms, $K_\lambda := \{x \in E : b_\lambda(x, \cdot) = 0\}$ and that

$$(i) \quad x \in K_\lambda \wedge a(x, \cdot)|_{K_\lambda} \Rightarrow x = 0$$

and for some $\alpha_\lambda, \beta_\lambda > 0$ there hold

$$(ii) \quad x \in K_\lambda \Rightarrow \alpha_\lambda \|x\| \leq \|a(\cdot, x)|_{K_\lambda}\|,$$

$$(iii) \quad y \in F \Rightarrow \beta_\lambda \|y\| \leq \|b(\cdot, y)\|.$$

If

$$\rho_\lambda := \max \left\{ \frac{1}{\alpha_\lambda}, \frac{1}{\beta_\lambda} \left(1 + \frac{\|a_\lambda\|}{\alpha_\lambda} \right), \frac{1}{\beta_\lambda^2} \|a_\lambda\| \left(1 + \frac{\|a_\lambda\|}{\alpha_j} \right) \right\}$$

and in addition

$$(iv) \quad \|c_\lambda\| < \frac{1}{\rho_\lambda},$$

then for each $\lambda \in \Lambda$ and $(x^*, y^*) \in E^* \times F^*$ there exists a unique $(x_\lambda, y_\lambda) \in E \times F$ such that

$$\begin{cases} a_\lambda(x_\lambda, \cdot) + b_\lambda(\cdot, y_\lambda) = x^* \\ b_\lambda(x_\lambda, \cdot) + c_\lambda(y_\lambda, \cdot) = y^* \end{cases} \quad (6)$$

Furthermore, if $(x, y) \in E \times F$, then

$$\max\{\|x_\lambda - x\|, \|y_\lambda - y\|\} \leq \frac{\rho_\lambda}{1 - \rho_\lambda \|c_\lambda\|} (\|x^* - a_\lambda(x, \cdot) - b_\lambda(\cdot, y)\| + \|y^* - b_\lambda(x, \cdot)\|). \quad (7)$$

The idea behind the inverse problem is that under the uniform conditions

$$\alpha := \inf_{\lambda \in \Lambda} \alpha_\lambda > 0, \quad \beta := \inf_{\lambda \in \Lambda} \beta_\lambda > 0, \quad \delta := \sup_{\lambda \in \Lambda} \|a_\lambda\|, \quad \gamma := \inf_{\lambda \in \Lambda} \|c_\lambda\| > 0$$

and

$$\rho := \max \left\{ \frac{1}{\alpha}, \frac{1}{\beta} \left(1 + \frac{\delta}{\alpha} \right), \frac{\delta}{\beta^2} \left(1 + \frac{\delta}{\alpha} \right) \right\},$$

then

$$\inf_{\lambda \in \Lambda} \max\{\|x_\lambda - x\|, \|y_\lambda - y\|\} \leq \frac{\rho}{1 - \rho\gamma} (\|x^* - a_\lambda(x, \cdot) - b_\lambda(\cdot, y)\| + \|y^* - b_\lambda(x, \cdot)\|).$$

So we can minimize

$$\{\|x^* - a_\lambda(x, \cdot) - b_\lambda(\cdot, y)\| : \lambda \in \Lambda\}$$

and

$$\{\|y^* - b_\lambda(x, \cdot)\| : \lambda \in \Lambda\},$$

and that is to solve the optimization problem

$$\min_{\lambda \in \Lambda} F(\lambda) := \|x_\lambda^* - a_\lambda(\hat{w}, \cdot) - b_\lambda(\cdot, \hat{\psi})\| + \|y_\lambda^* - b_\lambda(\hat{w}, \cdot)\|$$

and then we approximate the solution of the inverse problem.

Example 1 As in [4] we consider de the boundary value problem:

$$\begin{cases} \Delta^2 \psi + \delta \psi = f \text{ in } \Omega \\ \psi|_{\Gamma} = 0 \\ \Delta \psi|_{\Gamma} = 0 \end{cases}, \tag{8}$$

where $\Omega = (0, 1)^2$, $\Gamma = \partial\Omega$, $\delta \in \mathbb{R}$ and $f \in H_0^1(\Omega)$. If one takes $w := -\Delta\psi$, then this problem is equivalent to

$$\begin{cases} w + \Delta\psi = 0 \text{ in } \Omega \\ -\Delta w + \delta \psi = f \text{ in } \Omega \\ \psi|_{\Gamma} = 0 \\ w|_{\Gamma} = 0 \end{cases}, \tag{9}$$

and by easy passages it can be written in this variational formulation (5): find $(w, \psi) \in E \times F$ such that

$$\begin{cases} v \in E \Rightarrow a(w, v) + b(v, \psi) = x^*(v) \\ \phi \in W \Rightarrow b(w, \phi) + c(\psi, \phi) = y^*(\phi) \end{cases}.$$

This system adopts the form of (6) with $\text{card}(\Lambda) = 1$, the real Hilbert spaces $E = F := H_0^1(\Omega)$, the continuous bilinear forms $a : E \times E \rightarrow \mathbb{R}$, $b : E \times F \rightarrow \mathbb{R}$ and $c : F \times F \rightarrow \mathbb{R}$ defined for each $w, v \in E$, and $\phi, \psi \in F$, as

$$a(w, v) := \langle w, v \rangle,$$

$$b(v, \psi) := -\langle \nabla v, \nabla \psi \rangle,$$

and

$$c(\psi, \phi) := -\delta \langle \psi, \phi \rangle,$$

and the continuous linear forms $x^* := 0 \in E^*$ and $y^* \in F^*$ given by

$$y^*(\phi) := -\langle f, \phi \rangle, \quad (\phi \in F).$$

To run a numerical simulation, we use the model presented in the previous example and set $\delta = -2$, and $f(x, y)$ the function in such a way that the solution $\psi(x, y)$ to the problem is $10^3[x(1-x)y(1-y)]^4$. We suppose there are no holes and the domain is solid. We solve the system in COMSOL. Then we sample the numerical solution on

a uniform grid of 9×9 interior points of $[0, 1]^2$. We feed the resulting representation of ψ and w into our generalized collage theorem machinery and, knowing $f(x, y)$, we recover $C[1], C[2], C[3]$ so that these representations are approximate solutions to the system

$$\begin{cases} C[1]\Delta\psi + C[2]w = 0, \\ -C[1]\Delta w + C[3]\psi = f(x, y). \end{cases}$$

True values are $C[1] = 1, C[2] = 1, C[3] = -2$. The results for two runs, one with 1% relative noise added, are below. The final number is the value of the generalized collage distance.

Noise	C[1]	C[2]	C[3]	Collage distance
0.00	0.99998	1.00058	-1.85635	0.00045
0.01	1.00316	1.00411	-3.14588	0.00171

4 Inverse Problems on Perforated Domains: A Computational Study

The concept of porous media is essential in many areas of applied sciences and engineering, including chemical engineering, civil engineering, petroleum engineering, aerospace engineering, soil science, geology, and material science. A given material is said to be porous or perforated when it is characterized by a partitioning of the total volume a solid portion and the holes.

When a differential equation is formulated over a porous medium, the term “porous” implies that the state equation is written only on the solid domain while boundary conditions should be imposed on the entire boundary including the boundary of the holes. Since the porosity in materials can assume different forms and appear in varying degrees, solving differential equations over porous media is often a complicated task. Examples of this are Stokes or Navier-Stokes equations that are usually written for the fluid part while the rocks play the role of “mathematical” holes.

Given a compact and convex set Ω , we denote by Ω_B the collection of all holes B_j . We also suppose that each hole $B_j \subset B(x_j, \varepsilon_j)$ where $B(x_j, \varepsilon_j)$ is a ball centered at x_j and with radius ε_j . We let $\varepsilon = \max_j \varepsilon_j$. The purpose of the analysis of problems on perforated domains is to analyze the stability of the inverse problem estimation’s results whenever $\varepsilon \rightarrow 0$ and the balls B_j become smaller and smaller.

Going back to the model presented in Sect. 3, given a solid set $\Omega = (0, 1)^2$, $\Gamma = \partial\Omega$, let us consider the perforated domain $\Omega_\varepsilon = \Omega \setminus \cup_{i=1}^n B_j$ where ε is the radius of the biggest ball which contains the biggest hole B_j . Let us denote by $\Gamma_\varepsilon = \partial\Omega_\varepsilon$, $\delta \in \mathbb{R}$, and $f \in H_0^1(\Omega_\varepsilon)$. Consider the porous version of the boundary value problem in Example 1:

$$\begin{cases} \Delta^2 \psi + \delta \psi = f \text{ in } \Omega_\epsilon \\ \psi|_{\Gamma_\epsilon} = 0 \\ \Delta \psi|_{\Gamma_\epsilon} = 0 \end{cases} \quad (10)$$

Then, multiplying its first equation by a test function $v \in H_0^1(\Omega_\epsilon)$, and integrating by part, we arrive at

$$\int_{\Omega_\epsilon} wv - \int_{\Omega_\epsilon} \nabla w \nabla v = 0.$$

On the other hand, when multiplying the second equation of (10) by a test function $\phi \in H_0^1(\Omega_\epsilon)$, and, proceeding as above, we write it as

$$- \int_{\Omega_\epsilon} \nabla w \nabla \phi - \delta \int_{\Omega_\epsilon} \psi \phi = - \int_{\Omega_\epsilon} f \phi.$$

Therefore, if we take the Hilbert spaces $E_\epsilon = F_\epsilon := H_0^1(\Omega_\epsilon)$, the continuous bilinear forms $a_\epsilon : E_\epsilon \times E_\epsilon \rightarrow \mathbb{R}$, $b_\epsilon : E_\epsilon \times F_\epsilon \rightarrow \mathbb{R}$ and $c_\epsilon : F_\epsilon \times F_\epsilon \rightarrow \mathbb{R}$ defined for each $w, v \in E_\epsilon$ and $\phi, \psi \in F_\epsilon$, as

$$a_\epsilon(w, v) := \int_{\Omega_\epsilon} wv,$$

$$b_\epsilon(v, \psi) := - \int_{\Omega_\epsilon} \nabla v \nabla \psi,$$

and

$$c_\epsilon(\psi, \phi) := -\delta \int_{\Omega_\epsilon} \psi \phi,$$

and the continuous linear forms $x^* \in E_\epsilon^*$ and $y^* \in F_\epsilon^*$ given by

$$x_\epsilon^*(v) := 0 \quad (v \in E_\epsilon)$$

and

$$y_\epsilon^*(\phi) := - \int_{\Omega_\epsilon} f \phi, \quad (\phi \in F_\epsilon),$$

then, if we proceed by doing the same passages we did in the solid domain case, the above model can be rewritten in the following form: Find $(w_\epsilon, \psi_\epsilon) \in E_\epsilon \times F_\epsilon$ such that

$$\begin{cases} v \in E_\epsilon \Rightarrow a_\epsilon(w_\epsilon, v) + b(v, \psi_\epsilon) = x_\epsilon^*(v) \\ \phi \in F_\epsilon \Rightarrow b(w_\epsilon, \phi) + c(\psi_\epsilon, \phi) = y_\epsilon^*(\phi) \end{cases}.$$

Here we take $E_\epsilon = F_\epsilon := H_0^1(\Omega_\epsilon)$, and the continuous linear forms $x_\epsilon^*(v) := \langle 0, v \rangle$ and $y_\epsilon^*(\phi) := \langle -f, \phi \rangle$. One is interested in analyzing the behaviour of

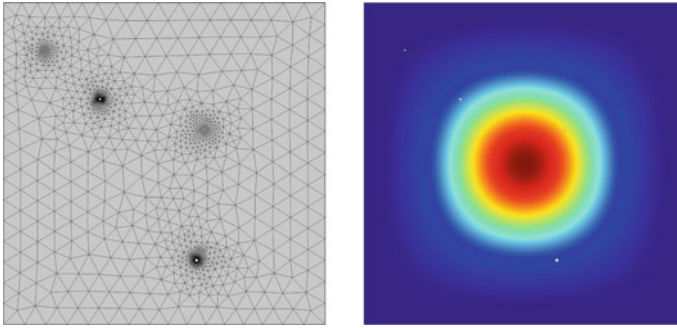


Fig. 1 Mesh and isotherms where the square side is 0.005, the circle radius is 0.005, the ellipse major axis is 0.005, and the minor axis is 0.003

the solution as well as the stability of the inverse problem results when $\epsilon \rightarrow 0$. This analysis is left to a future paper that will discuss these aspects in details. In the following three examples we run the collage coding approach over perforated domains in which the holes take different shapes and sizes. We randomly put four holes, and the shapes we use are squares, circles, and ellipses.

Example 2 In this example we consider the mesh and the isotherms are given in Fig. 1. In this example we suppose that the square side is 0.005, the circle radius is 0.005, the ellipse major axis is 0.005, and the minor axis is 0.003. The results are shown in the following table.

Elements	C[1]	C[2]	C[3]	Collage distance
05	1.000157776	1.002858381	-3.535085287	$3.427295666 * 10^{-12}$
10	1.000192021	1.003093897	-3.670642019	$1.964405817 * 10^{-12}$
15	1.000227419	1.003165604	-3.775473780	$1.081577887 * 10^{-12}$
20	1.000249733	1.003200522	-3.841718970	$6.683393914 * 10^{-13}$

Example 3 In this example we consider the mesh and the isotherms given in Fig. 2. In this example we set that the square side is 0.0005, the circle radius is 0.0005, the ellipse major axis is 0.0005, and the minor axis 0.0003. Results are shown in the following table.

Elements	C[1]	C[2]	C[3]	Collage distance
05	1.000155327	1.000282807	-2.095981040	$1.536328957 * 10^{-15}$
10	1.000157458	1.000297366	-2.104443730	$3.957312313 * 10^{-15}$
15	1.0001596550	1.0003018000	-2.1109728600	$2.303236318 * 10^{-15}$
20	1.000161039	1.000303959	-2.115096071	$1.533338673 * 10^{-15}$

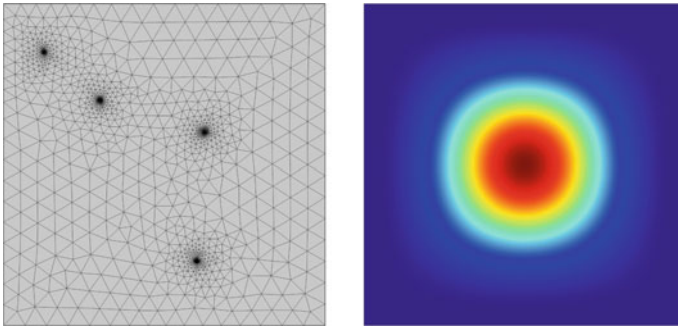


Fig. 2 Mesh and isotherms where the square side is 0.0005, the circle radius is 0.0005, the ellipse major axis is 0.0005, and the minor axis 0.0003

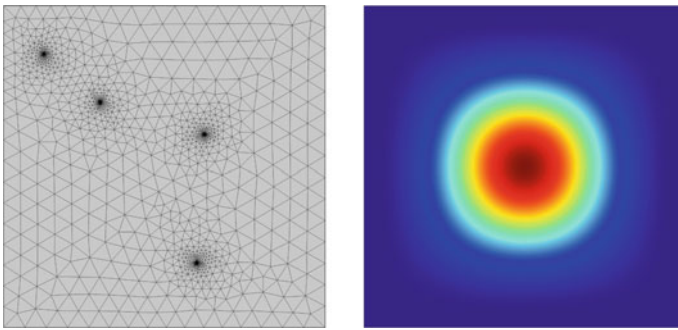


Fig. 3 Mesh and isotherms where the square side is 0.00005, the circle radius is 0.00005, the ellipse major axis is 0.00005, and the minor axis 0.00003

Example 4 In this example we consider the mesh and thhe isotherms are shown in Fig. 3. We suppose that the square side is 0.00005, the circle radius is 0.00005, the ellipse major axis is 0.00005, and the minor axis 0.00003. The results are shown in the following table.

Elements	C[1]	C[2]	C[3]	Collage distance
05	1.000154670	1.000255958	-2.080640364	$1.744800741 * 10^{-15}$
10	1.000156459	1.000268181	-2.087745875	$2.088385461 * 10^{-15}$
15	1.000158303	1.000271903	-2.093227440	$1.224364804 * 10^{-15}$
20	1.000159465	1.000273715	-2.096689098	$8.256778945 * 10^{-16}$

5 Conclusion

We have collected some results and an example in [4] related to a perturbed version of a mixed variational problem. We also have run several numerical examples to show how the collage coding works when the domain is perforated. The results show that the as hole diameter decreases, results improve. Moreover, with fixed diameter, results generally get better as n increases. There are n^2 finite element basis functions used in our collage distance calculation. This preliminary computational study suggests that further investigation is necessary when studying the stability of the inverse problem solution. This will be investigated in a future paper.

Acknowledgements Research partially supported by project MTM2016-80676-P (AEI/FEDER, UE) and by Junta de Andalucía Grant FQM359.

References

1. Barnsley, M.: *Fractals Everywhere*. Academic Press, New York (1989)
2. Berenguer, M.I., Kunze, H., La Torre, D., Ruiz Galán, M.: Galerkin method for constrained variational equations and a collage-based approach to related inverse problems. *J. Comput. Appl. Math.* **292**, 67–75 (2016)
3. Boffi, D. et al.: *Mixed finite elements, compatibility conditions and applications*. Lecture Notes in Math, vol. 1939. Springer-Verlag, Berlin (2008)
4. Garralda-Guillem, A.I., Kunze, H., La Torre, D., Ruiz Galán, M.: Using the Generalized Collage Theorem for Estimating Unknown Parameters in Perturbed Mixed Variational Equations. Submitted for publication
5. Garralda-Guillem, A.I., Ruiz Galán, M.: A minimax approach for the study of systems of variational equations and related Galerkin schemes. *J. Comput. Appl. Math.* **354**, 103–111 (2019)
6. Garralda-Guillem, A.I., Ruiz Galán, M.: Mixed variational formulations in locally convex spaces. *J. Math. Anal. Appl.* **414**, 825–849 (2014)
7. Hadamard, J.: *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. Yale University Press (1923)
8. Kirsch, A.: *An Introduction to the Mathematical Theory of Inverse Problems*. Springer (2011)
9. Kunze, H., Hicken, J., Vrscay, E.R.: Inverse problems for ODEs using contraction maps: suboptimality of the "collage method". *Inverse Prob.* **20**, 977–991 (2004)
10. Kunze, H., La Torre, D.: An inverse problem for a system of steady-state reaction-diffusion equations on a porous domain using a collage-based approach. *Journal of Physics: Conf. Ser.* **1047** (2018)
11. Kunze, H., Vrscay, E.R.: Solving inverse problems for ordinary differential equations using the Picard contraction mapping. *Inverse Prob.* **15**, 745–770 (1999)
12. Kunze, H., La Torre, D., Vrscay, E.R.: A generalized collage method based upon the Lax-Milgram functional for solving boundary value inverse problems. *Nonlinear Anal.* **71**, 1337–1343 (2009)
13. Kunze, H., La Torre, D., Vrscay, E.R.: Solving inverse problems for variational equations using the "generalized collage methods," with applications to boundary value problems. *Nonlinear Anal. Real World Appl.* **11**, 3734–3743 (2010)
14. Kunze, H., La Torre, D., Mendivil, F., Vrscay, E.R.: *Fractal-Based Methods in Analysis*. Springer (2012)

15. Kunze, H., La Torre, D.: Collage-type approach to inverse problems for elliptic PDEs on perforated domains. *Electron. J. Differ. Equ.* **48** (2015)
16. Kunze, H., La Torre, D.: An inverse problem for a 2-D system of steady-state reaction-diffusion equations on a perforated domain. *AIP Conf. Proc.* **1798** (2017)
17. Moura Neto, F.D., da Silva Neto, A.J.: *An Introduction to Inverse Problems with Applications*. Springer, New York (2013)
18. Tarantola, A.: *Inverse Problem Theory and Methods for Model Parameter Estimation* SIAM. Philadelphia (2005)
19. Tychonoff, A.N.: Solution of incorrectly formulated problems and the regularization method. *Doklady Akademii Nauk SSSR* **151**, 501–504 (1963)
20. Tychonoff, A.N., Arsenin, N.Y.: *Solution of Ill-posed Problems*. Winston & Sons, Washington (1977)
21. Vogel, C.R.: *Computational Methods for Inverse Problems*. SIAM, New York (2002)

bacoli_py—A Python Package for the Error Controlled Numerical Solution of 1D Time-Dependent PDEs



Connor Tannahill and Paul Muir

Abstract This paper introduces, *bacoli_py*, a Python 3 package for computing error controlled numerical solutions to 1D time-dependent PDEs. This package wraps modified versions of the Fortran packages, BACOLI and BACOLRI, so that they can be used in the more widely accessed and user-friendly Python 3 environment. This paper first provides an overview of the underlying numerical algorithms that are implemented in this package, followed by a description of the components of the package and then two examples to demonstrate its usage.

Keywords Partial differential equations · Error control · Python

1 Introduction

In this paper, we introduce *bacoli_py*, an open-source Python 3 module which can be used to compute *error controlled* numerical solutions to one dimensional time-dependent PDEs of the form

$$\mathbf{u}_t(t, x) = \mathbf{f}(t, x, \mathbf{u}(t, x), \mathbf{u}_x(t, x), \mathbf{u}_{xx}(t, x)), \quad x \in [x_a, x_b], \quad t \in [t_0, t_{out}], \quad (1)$$

where $\mathbf{u} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^n$ and $\mathbf{f} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, with initial conditions,

$$\mathbf{u}(t, x_0) = \mathbf{u}_0(x), \quad x \in [x_a, x_b], \quad (2)$$

where $\mathbf{u}_0 : \mathbb{R} \rightarrow \mathbb{R}^n$, and separated boundary conditions,

$$\mathbf{b}_L(t, \mathbf{u}(x_a, t), \mathbf{u}_x(x_a, t)) = \mathbf{0}, \quad \mathbf{b}_R(t, \mathbf{u}(x_b, t), \mathbf{u}_x(x_b, t)) = \mathbf{0}, \quad t \in [t_0, t_{out}], \quad (3)$$

This work was supported by the Natural Sciences and Engineering Research Council of Canada and Saint Mary's University.

C. Tannahill · P. Muir (✉)
Saint Mary's University, Halifax, NS B3H 3C3, Canada
e-mail: muir.smu@gmail.com

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_27

where $\mathbf{b}_L, \mathbf{b}_R : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\mathbf{0} \in \mathbb{R}^n$, and where $n \equiv npde$ is the number of PDEs.

Error control algorithms attempt to generate approximate solutions for which an associated high-quality error estimate satisfies a user-prescribed tolerance. In this way, the user can be reasonably confident that the approximate solution that is returned will have an error that approximately satisfies the tolerance. Additionally, the cost of the computation can be expected to be proportional to the tolerance.

The *bacoli_py* solver wraps the Fortran packages BACOLI [10] and BACOLRI [1], the most recent members of a family of B-spline Gaussian collocation error control solvers for this problem class. These solvers have been shown to efficiently compute error controlled numerical solutions to 1D PDEs of the form (1)–(3) [11, 12]. Prior to being employed within this Python module, these solvers were modified in two significant ways: (i) the linear system solver COLROW [3] was replaced with the LAMPAK solver [6] (to ensure copyright compliance within the Python 3 environment) and (ii) the calls to the function that defines the right hand side of (1) were modified such that the cross-language callbacks could be done more efficiently through vectorization using *numpy* [7] arrays. The purpose of this vectorization is to minimize the number of cross-language callbacks used by re-organizing the code such that repeated evaluations of the main user callback routine are instead done in one larger call to the modified user routine. We have also developed, as an alternative, an option that allows *bacoli_py* to make use of compiled Fortran subroutines. The Python to Fortran interface required for this package is generated using *f2py* [8].

Having these solvers available within the Python language will greatly increase the potential user community compared to that of the corresponding Fortran solvers. Furthermore, because of the capabilities of the Python language, substantial simplifications can be made in the user interface compared to what is necessary when using the lower-level Fortran codes directly. As well, users can take advantage of the many high-quality tools which are easily available within the Python ecosystem for analysis and visualization of results.

These advantages do come at a cost to performance, with *bacoli_py*, in its standard mode, i.e., with vectorized calls to Python callback functions, being at least an order of magnitude slower than the Fortran solvers, BACOLI and BACOLRI. The Python module is therefore primarily useful for initial prototyping and model exploration. For applications where efficiency is of major concern, the authors recommend the use of either the Fortran versions of this software, available at <http://cs.smu.ca/~muir/BACOLI-3Webpage.htm>, or the use of *bacoli_py* with callback functions as compiled Fortran subroutines; both of these options are at least an order of magnitude faster than *bacoli_py*, in standard mode.

For complete documentation, see <https://bacoli-py.readthedocs.io/en/latest/>. The *bacoli_py* package is available from PyPi at <http://pypi.python.org/pypi/bacoli-py>. The source code is available at https://github.com/connortannahill/bacoli_py.

2 Overview of BACOLI and BACOLRI

The BACOLI and BACOLRI packages represent the approximate solution to (1)–(3) at a given point in time, t , as a linear combination of C^1 -continuous B-spline basis functions [2] of a given degree p . Let $x_a = x_0 < x_1 < \dots < x_{nint} = x_b$ be a mesh with $nint$ subintervals that partitions the spatial domain $[x_a, x_b]$. Then the approximate solution is represented as

$$\mathbf{U}(t, x) = \sum_{i=1}^{NC_p} \mathbf{y}_{p,i}(t) B_{p,i}(x), \quad (4)$$

where $\mathbf{y}_{p,i}(t)$ is the unknown time-dependent vector coefficient of $B_{p,i}(x)$, the i th B-spline basis function of degree p , and $NC_p = nint(p - 1) + 2$. Equations for the determination of the unknown coefficients in (4), $\mathbf{y}_{p,i}(t)$, are obtained by requiring that (4) exactly satisfies (1) at $p - 1$ collocation points on each subinterval. These conditions have the form

$$\mathbf{U}_l(t, \xi_l) = \mathbf{f}(t, \xi_l, \mathbf{U}(t, \xi_l), \mathbf{U}_x(t, \xi_l), \mathbf{U}_{xx}(t, \xi_l)), \quad (5)$$

for $l = 2, \dots, NC_p - 1$, where the collocation points are $\xi_l = x_{i-1} + h_i \rho_j$, for $l = 1 + (i - 1)(p - 1) + j$, $i = 1, \dots, nint$, $j = 1, \dots, p - 1$, $\{\rho_i\}_{i=1}^{p-1}$ are the images of the order $p - 1$ Gauss points on $[0, 1]$, and $h_i = x_i - x_{i-1}$. As well, at $\xi_1 = x_a$, $\xi_{NC_p} = x_b$, (4) is required to satisfy the BCs, giving,

$$\mathbf{b}_L(t, \mathbf{U}(t, x_a), \mathbf{U}_x(t, x_a)) = \mathbf{0}, \quad \mathbf{b}_R(t, \mathbf{U}(t, x_b), \mathbf{U}_x(t, x_b)) = \mathbf{0}. \quad (6)$$

The system of time-dependent ordinary differential equations, (5), coupled with the conditions, (6), forms a system of Differential Algebraic Equations (DAEs) which is solved for $\mathbf{y}_{p,i}(t)$ using standard error control solvers for DAEs. BACOLI uses DASSL [9] for solving (5)–(6), whereas BACOLRI uses RADAU5 [5]. DASSL makes use of a family of multi-step methods called Backwards Differentiation Formulas (BDFs). RADAU5 is based on a fifth order Implicit Runge Kutta (IRK) method of Radau IIA type. The resultant approximate solution (4) has a spatial error that is $\mathcal{O}(h^{p+1})$, where h is the maximum spatial mesh subinterval size; see, e.g., [10] and references within.

On standard test problems, the two codes have comparable performance but BACOLRI out-performs BACOLI for certain classes of problems where stability issues arise for the higher-order BDFs. Such problems are characterized as those which lead to DAE systems that have Jacobians with eigenvalues near the imaginary axis. It is well known that the stability regions of the higher order BDFs have gaps near the imaginary axis, just above and below the origin, and thus are not stable for such problems. See [12] and references within.

After (4) has been obtained for a given point in time, a spatial error estimate is computed. This error estimate requires the computation of a second approximate

solution, $\bar{\mathbf{U}}(t, x)$, which has spatial error of a different order of accuracy than that of $\mathbf{U}(t, x)$. A scaled difference of $\mathbf{U}(t, x)$ and $\bar{\mathbf{U}}(t, x)$ then gives the spatial error estimate. The approximate solution, $\mathbf{U}(t, x)$, at the current time step is accepted if the error estimate meets the tolerance. Otherwise, the step is rejected and a spatial remeshing algorithm is applied which attempts to compute a new spatial mesh such that the computed solution obtained on this new mesh will have a spatial error estimate that satisfies the tolerance. This remeshing algorithm works by (i) adjusting the number of mesh subintervals based on the magnitude of the spatial error estimate and (ii) using equidistribution to re-position the mesh points into regions where the spatial error estimate is largest.

BACOLI and BACOLRI make use of inexpensive, interpolation-based spatial error estimates referred to as the SuperConvergent Interpolation (SCI) scheme and the Lower Order Interpolation (LOI) scheme. For each spatial subinterval, the SCI interpolates $\mathbf{U}(t, x)$ or $\mathbf{U}_x(t, x)$ at known points of superconvergence where these values have higher orders of accuracy than at arbitrary points within $[x_a, x_b]$. In this case, $\bar{\mathbf{U}}(t, x)$ is a C^1 -continuous piecewise polynomial based on Hermite-Birkhoff interpolating polynomials on each subinterval which interpolate a set of superconvergent values associated with the subinterval. The scaled difference of $\mathbf{U}(t, x)$ and $\bar{\mathbf{U}}(t, x)$ provides an estimate of the error of $\mathbf{U}(t, x)$ and this is referred to as Standard (ST) error control. The LOI scheme implements an alternative form of error control known as Local Extrapolation (LE) error control. In LE error control, an approximate solution, $\bar{\mathbf{U}}(t, x)$, of one spatial order of accuracy less than that of $\mathbf{U}(t, x)$ is computed. The scaled difference of $\mathbf{U}(t, x)$ and $\bar{\mathbf{U}}(t, x)$ in this case gives an estimate of the spatial error of $\bar{\mathbf{U}}(t, x)$, which provides a conservative upper bound on the error in $\mathbf{U}(t, x)$. In the LOI scheme, this lower order approximation is expressed as a Hermite-Birkhoff interpolant on each subinterval which has been constructed such that its interpolation error is asymptotically equivalent to the leading order error term in a collocation solution of one order of accuracy less than $\mathbf{U}(t, x)$. See [10] for further details. The SCI and LOI schemes thus provide *bacoli_py* with two types of error control, ST error control or LE error control. This is similar to what is provided with Runge-Kutta formula pairs for the numerical solution of initial value ODEs; see, e.g., [4].

For more complete descriptions of the BACOLI/BACOLRI algorithms, see [10, 12] and references within.

3 Description of *bacoli_py*

3.1 Basic Usage

bacoli_py provides a convenient, minimal, object-oriented programming interface that is substantially simpler than that of the Fortran packages. For standard usage of *bacoli_py* the user must first define the system of *npde* PDEs to be solved in terms

of the Python callback functions, f , $bndxa$, $bndxb$, $uinit$, which correspond to (1), (3), and (2), respectively. These are encapsulated within a *ProblemDefinition* object as

```
| problem_definition =
|     bacoli_py.ProblemDefinition(npde, f, bndxa, bndxb, uinit).
```

A *Solver* object which performs the main functionality of *bacoli_py* is then initialized. This can be done simply by

```
| solver = bacoli_py.Solver().
```

The *Solver* object contains the method, *solve*, which is used to solve (1)–(3) defined by a *ProblemDefinition* object. The arguments to this method include a *ProblemDefinition* object, the initial time, t_0 , the spatial boundaries, x_a , x_b , and the points in time and space at which the solution values are required. A call to *solve* takes the form

```
| evaluation = solver.solve(problem_definition,
|                           initial_time, [xa,xb], tspan, xspan).
```

This call returns an *Evaluation* object containing the computed solution information. In particular, the approximate solution is contained, as an attribute in this object, in a *numpy* array with dimensions $(npde, len(tspan), len(xspan))$,

```
| u = evaluation.u
```

This summarizes the process of using *bacoli_py* in the majority of use cases. For a more complete description of this module, including its overall structure and descriptions of the many arguments and settings that can be specified, we refer the reader to <https://bacoli-py.readthedocs.io/en/latest/>.

4 Examples

In this section, we provide examples in which *bacoli_py* is applied to two test problems. See <https://bacoli-py.readthedocs.io/en/latest/> for additional examples.

4.1 One Layer Burgers Equation—Python Callback Functions

The One Layer Burgers Equation is given by,

$$u_t(t, x) = \epsilon u_{xx}(t, x) - u(t, x)u_x(t, x), \quad x \in [0, 1], \quad t \in [0, 1],$$

where the initial conditions and the Dirichlet boundary conditions taken from the exact solution,

$$u(t, x) = \frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x - \frac{t}{2} - \frac{1}{4}}{4\epsilon}\right),$$

and where ϵ is chosen to be 10^{-3} . We choose a tolerance of 10^{-6} .

We first describe this system in terms of Python callback functions, as well as globally defining *npde* and the problem-dependent parameter ϵ . These functions are then placed within a *ProblemDefinition* object.

```
import bacoli_py
import numpy
from numpy import tanh, array
# Specify the number of PDEs in this system.
npde = 1

# Initialize problem-dependent parameter.
eps = 1.0e-3

# Function defining the PDE.
def f(t, x, u, ux, uxx, fval):
    fval[0] = eps*uxx[0] - u[0]*ux[0]
    return fval

# Function defining the left spatial boundary condition.
def bndxa(t, u, ux, bval):
    bval[0] = u[0] - 0.5 + 0.5*tanh( (-0.5*t-0.25) / (4.0*eps) )
    return bval

# Function defining the right spatial boundary condition.
def bndxb(t, u, ux, bval):
    bval[0] = 0.5*tanh((0.75-0.5*t)/(4.0*eps)) - 0.5 + u[0]
    return bval

# Function defining the initial condition.
def uinit(x, u):
    u[0] = 0.5 - 0.5 * tanh((x - 0.25) / (4.0*eps))
    return u

# Pack all of these callbacks and the number of PDEs into a
# ProblemDefinition object.
problem_definition = bacoli_py.ProblemDefinition(npde, f=f,
                                                bndxa=bndxa, bndxb=bndxb, uinit=uinit)
```

Once *ProblemDefinition* is created, using *bacoli_py* to solve (1)–(3) is fairly straightforward, requiring only the creation of a *Solver* object and the specification of

- The initial time t_0 ,
- An array, $[x_a, x_b]$, containing the spatial boundary points,
- The points at which the solution will be evaluated. (This is done by providing a list of x points, $xspan$, and t points, $tspan$, at which the evaluations of the numerical solution will be provided.)
- The absolute ($atol$) and relative ($rtol$) error tolerances. (Here we set $atol = rtol = 10^{-6}$.)

```
# Initialize the Solver object.
solver = bacoli_py.Solver()

# Set t0.
initial_time = 0.0

# Define the spatial boundaries.
initial_mesh = numpy.array([0.0, 1.0])
# Choose output times and points. Here our final time t_end = 1.
tspan = numpy.linspace(0.001, 1, 100)
xspan = numpy.linspace(0, 1, 100)

# Solve this problem.
evaluation = solver.solve(problem_definition, initial_time,
    initial_mesh, tspan, xspan, atol=1e-6, rtol=1e-6, dirichlet=True)
```

The solution obtained from this call is plotted in Fig. 1.

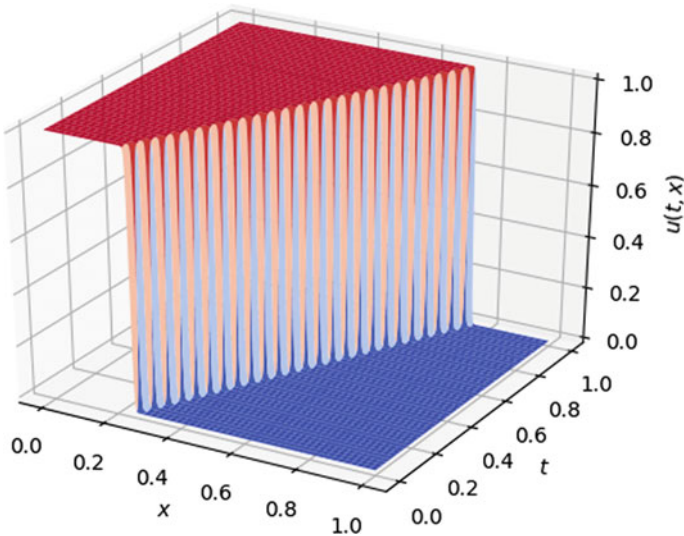


Fig. 1 One layer burgers equation, $\epsilon = 10^{-3}$

4.2 Two Layer Burgers Equation—Callback Functions as Compiled Fortran Subroutines

To increase the efficiency of *bacoli_py*, it is possible to define the callback functions in a *ProblemDefinition* as compiled Fortran subroutines. In this example, we demonstrate how this can be done in the context of solving the Two Layer Burgers Equation. This PDE is given by

$$u_t(t, x) = \epsilon u_{xx}(t, x) - u(t, x)u_x(t, x), \quad x \in [0, 1], \quad t \in [0, 1],$$

with initial and (Dirichlet) boundary conditions taken from the exact solution,

$$u(t, x) = \frac{0.1e^{-A} + 0.5e^{-B} + e^{-C}}{e^{-A} + e^{-B} + e^{-C}}, \quad t \in [0, 1], \quad x \in [0, 1],$$

where,

$$A = \frac{0.05}{\epsilon}(x - 0.5 + 4.95t), \quad B = \frac{0.25}{\epsilon}(x - 0.5 + 0.75t), \quad C = \frac{0.5}{\epsilon}(x - 0.375),$$

and where ϵ is chosen to be 10^{-4} . We use the default tolerance, 10^{-4} .

We first define Fortran 95 callback routines for each of the callback functions representing the PDE, its initial condition, and its boundary conditions. The *numpy.f2py* module is used to build an extension module containing these Fortran subroutines, callable from Python.

```
import numpy.f2py as f2py

# String defining the Fortran 95 callback subroutines.
prob_def_f = """
    subroutine f(t, x, u, ux, uxx, fval)
        integer          npde
        parameter        (npde=1)
        double precision t, x, u(npde), ux(npde)
        double precision uxx(npde), fval(npde)
        double precision eps
        parameter        (eps=1d-4)
        fval(1) = eps*uxx(1) - u(1)*ux(1)
    return
end
subroutine bndxa(t, u, ux, bval)
    integer          npde
    parameter        (npde=1)
    double precision t, u(npde), ux(npde), bval(npde)
    double precision eps
```

```

parameter      (eps=1d-4)
double precision a1, a2, a3, expa1, expa2, expa3, temp
a1 = (0.5d0 - 4.95d0 * t) * 0.5d-1 / eps
a2 = (0.5d0 - 0.75d0 * t) * 0.25d0 / eps
a3 = 0.1875d0 / eps
expa1 = 0.d0; expa2 = 0.d0; expa3 = 0.d0
temp = max(a1, a2, a3)
if ((a1-temp) .ge. -35.d0) expa1 = exp(a1-temp)
if ((a2-temp) .ge. -35.d0) expa2 = exp(a2-temp)
if ((a3-temp) .ge. -35.d0) expa3 = exp(a3-temp)
bval(1) = u(1) - (0.1d0*expa1+0.5d0*expa2+expa3) &
           / (expa1+expa2+expa3)

return
end

subroutine bndxb(t, u, ux, bval)
integer      npde
parameter    (npde=1)
double precision t, u(npde), ux(npde), bval(npde)
double precision eps
parameter    (eps=1d-4)
double precision a1, a2, a3, expa1, expa2, expa3, temp
a1 = (-0.5d0 - 4.95d0 * t) * 0.5d-1 / eps
a2 = (-0.5d0 - 0.75d0 * t) * 0.25d0 / eps
a3 = - 0.3125d0 / eps
expa1 = 0.d0; expa2 = 0.d0; expa3 = 0.d0
temp = max(a1, a2, a3)
if ((a1-temp) .ge. -35.d0) expa1 = exp(a1-temp)
if ((a2-temp) .ge. -35.d0) expa2 = exp(a2-temp)
if ((a3-temp) .ge. -35.d0) expa3 = exp(a3-temp)
bval(1) = u(1) - (0.1d0*expa1+0.5d0*expa2+expa3) &
           / (expa1+expa2+expa3)

return
end

subroutine uinit(x, u)
integer      npde
parameter    (npde=1)
double precision x, u(npde)
double precision eps
parameter    (eps=1d-4)
double precision a1, a2, a3, expa1, expa2, expa3, temp
a1 = (-x + 0.5d0) * 0.5d-1 / eps
a2 = (-x + 0.5d0) * 0.25d0 / eps
a3 = (-x + 0.375d0) * 0.5 / eps

```

```

        expa1 = 0.d0; expa2 = 0.d0; expa3 = 0.d0
        temp = max(a1, a2, a3)
        if ((a1-temp) .ge. -35.d0) expa1 = exp(a1-temp)
        if ((a2-temp) .ge. -35.d0) expa2 = exp(a2-temp)
        if ((a3-temp) .ge. -35.d0) expa3 = exp(a3-temp)
        u(1) = (0.1d0*expa1+0.5d0*expa2+expa3) &
                /(expa1+expa2+expa3)

    return
end
"""

# Build extension module containing these callbacks.
f2py.compile(prob_def_f, modulename='problemdef', verbose=0,
              extension='.f95')

```

After the extension module has been built, the compiled callback functions can be used with *bacoli_py*. To do this, a *ProblemDefinition* object is defined which contains pointers to these compiled subroutines in place of the usual Python functions.

```

import bacoli_py
import numpy

# Import Fortran callbacks from extension module.
from problemdef import f, bndxa, bndxb, uinit

# Specify the number of PDEs in this system.
npde = 1

# Pack all of these callbacks and the number of PDEs into a
# ProblemDefinition object.
problem_definition = bacoli_py.ProblemDefinition(npde,
        f=f._cpointer, bndxa=bndxa._cpointer, bndxb=bndxb._cpointer,
        uinit=uinit._cpointer)

```

bacoli_py can then be used in almost entirely the same way as we saw in the previous example. The only exception to this is that the flag *compiled_callbacks* in the call to the *Solver.solve* method must be set to *True*, in which case the usual vector optimizations used in *bacoli_py* will not be employed.

```

# Initialize the Solver object.
solver = bacoli_py.Solver()

# Set t0.
initial_time = 0.0

# Define the spatial boundaries.
initial_mesh = [0, 1]

```

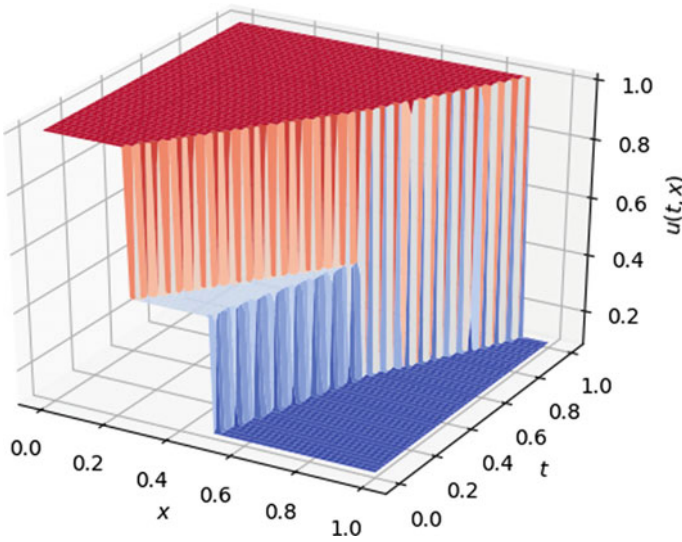


Fig. 2 Two layer burgers equation, $\epsilon = 10^{-4}$

```
# Choose output times and points.
tspan = numpy.linspace(0.001, 1, 100)
xspan = numpy.linspace(0, 1, 100)

# Solve this problem.
evaluation = solver.solve(problem_definition, initial_time,
                           initial_mesh, tspan, xspan, compiled_callbacks=True, dirichlet=True)
```

The computed solution is plotted in Fig. 2.

References

1. Adams, M., Tannahill, C., Muir, P.: Error control Gaussian collocation software for boundary value ODEs and 1D time-dependent PDEs. *Numer. Algor.* **81**, 1505–1519 (2019)
2. De Boor, C.: *A Practical Guide to Splines*, vol. 27. Springer-Verlag, New York (1978)
3. Díaz, J., Fairweather, G., Keast, P.: Algorithm 603: COLROW and ARCECO: FORTRAN packages for solving certain almost block diagonal linear systems by modified alternate row and column elimination. *ACM Trans. Math. Softw.* **9**, 376–380 (1983)
4. Hairer, E., Nørsett, S.P., Wanner, G.: *Solving Ordinary Differential Equations I*, volume 8 of Springer Series in Computational Mathematics, 2nd edn. Springer-Verlag, Berlin (1993)
5. Hairer, E., Wanner, G.: *Solving Ordinary Differential Equations II*. Springer Series in Computational Mathematics, vol. 14. Springer-Verlag, Berlin (1988)
6. Keast, P.: LAMPAK: a Fortran package for solving certain almost block diagonal matrices. Unpublished Software
7. Oliphant, T.: *NumPy: a guide to NumPy*. Trelgol Publishing, USA (2006) [Online; accessed June 5, 2019]

8. Peterson, P.: F2py: a tool for connecting Fortran and Python programs. *Int. J. Comp. Sci. Eng.* **4**, 296–305 (2009)
9. Petzold, L.: Description of DASSL: a differential/algebraic system solver. Technical report, Sandia National Labs., Livermore, CA, USA (1982)
10. Pew, J., Li, Z., Muir, P.: Algorithm 962: BACOLI: B-spline adaptive collocation software for PDEs with interpolation-based spatial error control. *ACM Trans. Math. Softw.* **42**, 25 (2016)
11. Pew, J., Li, Z., Tannahill, C., Muir, P., Fairweather, G.: Performance analysis of error-control B-spline Gaussian collocation software for PDEs. *Comput. Math. Appl.* **77**, 1888–1901 (2019)
12. Pew, J., Murtha, T., Tannahill, C., Muir, P.: Error control B-spline Gaussian collocation/Runge-Kutta PDE software with interpolation-based spatial error estimation. Technical Report 2018_002, Department of Mathematical and Computational Science Technical Report Series (2018)

Solving Cardiac Bidomain Problems with B-spline Adaptive Collocation



Kevin R. Green and Raymond J. Spiteri

Abstract B-spline collocation methods have been shown to be effective for solving systems of parabolic partial differential equations (PDEs). Using B-spline bases for spatial discretization and backward differentiation formulae for temporal discretization, software can be developed that allows full spatio-temporal error control throughout the solution. The software package `eBACOLI`, which uses \mathcal{C}^1 -continuous B-splines, has been extended for this approach to work with PDEs in one spatial dimension that have a multi-scale structure like the bidomain model, i.e., parabolic and elliptic PDEs at the macro-scale coupled with ordinary differential equations at the micro-scale. We present numerical results of cardiac bidomain simulations, validating them through comparison with solutions obtained from the software package `Nektar++`. The performance of `eBACOLI` and `Nektar++` simulations are compared by considering solution times with comparable error with respect to a reference solution, showing that in addition to automatically controlling the error (a feature unavailable in `Nektar++`), `eBACOLI` is generally more than an order of magnitude faster than `Nektar++` for a given error level.

Keywords Bidomain model · B-splines · Adaptive collocation

1 Introduction

The bidomain model is a continuum model used for simulating the electrical activity of cardiac tissue. It consists of a system of multi-scale partial differential equations (PDEs) that couples a signal propagation model in the tissue at the macro-scale with cardiomyocyte firing models at the micro-scale. The macro-scale model describes both the intra- and extra-cellular electrical potentials of the tissue sepa-

K. R. Green (✉) · R. J. Spiteri

Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada
e-mail: kevin.green@usask.ca

R. J. Spiteri

e-mail: spiteri@cs.usask.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_28

301

rately while being co-located in the domain of interest. The bidomain model was originally derived by Tung in the 1970s [21] and has been studied extensively since; see, e.g., [6, 9, 10, 13, 20] and references therein. Here, we use the more commonly written *Roth form* of the bidomain model [18],

$$\chi C_m \frac{\partial v}{\partial t} = \chi I_{\text{ion}}(\mathbf{x}, t, v, \mathbf{s}) + I_{\text{stim}}(\mathbf{x}, t) + \nabla \cdot (\sigma_i \nabla v) + \nabla \cdot (\sigma_e \nabla u_e), \quad (1a)$$

$$\frac{\partial \mathbf{s}}{\partial t} = \mathbf{f}(\mathbf{x}, t, v, \mathbf{s}), \quad (1b)$$

$$0 = \nabla \cdot (\sigma_i \nabla v) + \nabla \cdot ((\sigma_i + \sigma_e) \nabla u_e). \quad (1c)$$

In this formulation, the dynamical quantities are the transmembrane potential $v(\mathbf{x}, t)$, the extracellular potential $u_e(\mathbf{x}, t)$, and the cell model state variables $\mathbf{s}(\mathbf{x}, t)$, each defined for spatial domain $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ of dimension d and temporal domain $t \in [t_0, t_f]$. A cell model consists of the (generally nonlinear) functions $\mathbf{f}(\cdot)$, modelling its own dynamics, and $I_{\text{ion}}(\cdot)$, modelling its coupling to the macro-scale tissue. Parameters of the bidomain model are the cell membrane capacitance per unit area C_m , the cell membrane area per unit volume χ , and possibly many others needed within the cell model functions. Finally, $\sigma_i(\mathbf{x})$ and $\sigma_e(\mathbf{x})$ are the intra- and extra-cellular conductivity tensors, respectively. An external stimulus applied to the tissue is denoted by $I_{\text{stim}}(\cdot)$. Boundary conditions for the bidomain model are based on the assumption that ions do not flow out of the domain,

$$\begin{aligned} \hat{\mathbf{n}} \cdot (\sigma_i \nabla v + \sigma_e \nabla u_e) &= 0, \\ \hat{\mathbf{n}} \cdot (\sigma_e \nabla u_e) &= 0, \end{aligned} \quad (2)$$

where $\hat{\mathbf{n}}$ is the unit normal pointing outwards on the boundary $\partial\Omega$.

BACOLI is a member of a family of software packages that solves one-dimensional parabolic PDEs using adaptive B-spline collocation [16]. The software uses \mathcal{C}^1 -continuous B-splines for spatial discretization and the variable order backward differentiation formulae (BDF) package DASSL [15] for adaptive temporal discretization. Full spatial error control is applied when DASSL fails to find a suitable timestep size. The spatial representation is then refined globally to equidistribute spatial error according to an error estimator that is built from spatial interpolants.

BACOLI has recently been extended to eBACOLI [8], which is capable of handling multi-scale equations that couple parabolic PDE systems to spatially localized ordinary differential equation (ODE) systems; i.e., the capability to couple Eqs. (1a) to (1b) was added. The current work describes a further extension that permits the solution of equations of the form of the bidomain model that also includes non-dynamical *constraint* equations like Eq. (1c). At present, the software is only capable of handling a single elliptic equation as is needed for solving Eq. (1c).

The remainder of this paper progresses as follows. A review of the formulation of discretization used in eBACOLI and its relevance to the bidomain model are

provided in Sect. 2. The numerical experiments performed are described in Sect. 3. The results of the numerical experiments are presented in Sect. 4. Conclusions and future directions are discussed in Sect. 5.

2 The Bidomain Model in eBACOLI

This section gives a summary of the basics of adaptive B-spline collocation methods as needed for understanding a bidomain solve with eBACOLI. To use eBACOLI for cardiac bidomain simulation, we consider Eq. (1) with $d = 1$ and pack all of the dynamical quantities into a single vector $\mathbf{y}(x, t) = [v(x, t) \mathbf{s}^T(x, t) u_e(x, t)]^T$.

2.1 B-spline Basis Expansion

eBACOLI solves problems like Eq. (1) in the space $\mathbb{R}^{m_{\text{PDE}}} \times \mathbb{P}_{p,\pi,2}$, where m_{PDE} is the total number of PDEs of the system and $\mathbb{P}_{p,\pi,v}$ is the space of piecewise polynomials of degree p on the mesh

$$\pi = \{x_i \mid x_L = x_0 < x_1 < \dots < x_N = x_R\}$$

that are \mathcal{C}^{v-1} -continuous at the internal mesh points. For the bidomain model, $m_{\text{PDE}} = n_s + 2$, where n_s is the number of dynamic state variables in the cell model.

eBACOLI uses polynomials of order $p = 3, 4, \dots, 11$. The lower bound is based on how eBACOLI constructs spatial interpolants for error estimation, and the upper bound is due to practical considerations, chosen as the point at which the accuracy of standard double-precision calculations typically saturates due to roundoff errors [8].

With the degree p and mesh π fixed, a B-spline basis $\{B_{p,j}\}_{j=1}^M$ for $\mathbb{P}_{p,\pi,2}$ can be computed using divided differences [3]. The space $\mathbb{P}_{p,\pi,2}$ has dimension $M = N(p - 1) + 2$, and thus the dimension of the spatially discretized system is $m_{\text{PDE}}M$.

The approximate B-spline solution to (1) can therefore be written as

$$\mathbf{Y}(x, t) := [V(x, t) \mathbf{S}^T(x, t) U_e(x, t)]^T = \sum_{j=1}^M \mathbf{y}_{p,j}(t) B_{p,j}(x) \approx \mathbf{y}(x, t), \quad (3)$$

where $\mathbf{y}_{p,j}(t)$ is the vector of coefficients of B-spline basis function j of degree p .

2.2 Collocation Equations

To account for all of the degrees of freedom in the B-spline representation of the bidomain solution, we require the following:

- $p - 1$ collocation points, ρ_k , $k = 1, 2, \dots, p - 1$, on each interval for v , \mathbf{s} , and u_e on the interior of the domain. These collocation points are taken to be the canonical Gauss points on $[-1, 1]$ mapped to each interval;
- n_s collocation points on each boundary for \mathbf{s} at the boundaries;
- Two boundary conditions on each boundary for v and u_e at the boundaries.

The collocation conditions on the interior of the domain can be expressed as the system of differential-algebraic equations (DAEs)

$$\frac{d}{dt} \begin{bmatrix} \chi C_m V(\xi_l, t) \\ \mathbf{S}(\xi_l, t) \\ 0 \end{bmatrix} = \begin{bmatrix} \chi I_{\text{ion}}(\xi_l, t, V, \mathbf{S}) + I_{\text{stim}}(\xi_l, t) + (\sigma_i V_x)_x + (\sigma_i U_{e,x})_x \\ \mathbf{f}(\xi_l, t, V, \mathbf{S}) \\ (\sigma_i V_x)_x + ((\sigma_i + \sigma_e) U_{e,x})_x \end{bmatrix}, \quad (4a)$$

where subscript x denotes differentiation with respect to x and all dynamical fields have argument (ξ_l, t) for the collocation points

$$\begin{aligned} \xi_l &= x_{i-1} + \frac{h_i}{2} (\rho_k + 1), \quad \text{where } \begin{aligned} &i = 1, 2, \dots, N, \\ &k = 1, 2, \dots, p - 1, \\ &l = 1 + (i - 1)(p - 1) + k. \end{aligned} \end{aligned}$$

The relationship of Eq. (3) combined with a known B-spline basis for a given mesh allows all spatial derivatives of the approximate fields to be computed using coefficients of the B-spline expansion. Details can be found in [3].

At each boundary point $x_B = \{x_L, x_R\}$, we have the DAE system

$$\frac{d}{dt} \begin{bmatrix} 0 \\ \mathbf{S}(x_B, t) \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma_i V_x(x_B, t) + \sigma_i U_{e,x}(x_B, t) \\ \mathbf{f}(x_B, t, V(x_B, t), \mathbf{S}(x_B, t)) \\ \sigma_e U_{e,x}(x_B, t) \end{bmatrix}. \quad (4b)$$

The complete system Eq. (4) constitutes an index-1 DAE and is thus suitable for solution via the DASSL package. The version of eBACOLI described in [8] cannot handle Eq. (4) due to the algebraic equations present at every interior collocation point and was modified to obtain the results reported here. The modifications performed are mainly in setting up the Jacobian and residual evaluations for DASSL to use at each timestep. The last component in Eq. (4b) is handled in a similar way to its first component. The last component of Eq. (4a) is handled similarly to a boundary condition in terms of its dynamical behaviour but at the same time with spatial dependence that is similar to its first component in Eq. (4a).

2.3 Error Control

The error control remains unchanged from the original BACOLI package [16]. To summarize, temporal error is controlled via DASSL by approximating the error present in the BDF approximation of the DAE system and adjusting the timestep size and order of the BDF method according to absolute and relative tolerances.

With a successful time step from DASSL, a spatial error estimate is formed by comparing the B-spline solution to a second solution obtained by a different interpolation method. The interpolated solution can be of lower degree (lower-order interpolation (LOI)) or higher degree (super-convergent interpolation (SCI)) depending on the solution parameter settings within eBACOLI . If the spatial error approximation does not meet the desired tolerance, a remeshing algorithm is applied that tries to (i) estimate the number of mesh points necessary to do so and (ii) approximately equidistribute the error on the new mesh. This spatial remeshing creates a new spatial discretization and thus a new system of DAEs to integrate in time.

3 Numerical Experiments

We consider a single case for validation of bidomain solutions that uses conductivities σ_e and σ_i that are constant in space and time. In such a case, Eq. (1) reduces to the so-called *monodomain model*. This construct allows us to compare bidomain solutions with equivalent monodomain solutions, which, for example, can be computed using the original version of eBACOLI [8] when $d = 1$. For brevity, we do not report details of these comparisons here, noting only successful validation of the new software for this case, and report only results from numerical experiments that solve the problem as a bidomain model.

The FitzHugh–Nagumo (FHN) cell model [7, 12] is chosen for our experiments, with its functions taking the form

$$\begin{aligned} f_1(\mathbf{x}, t, v, s_1) &= \epsilon (v + \beta - \gamma s_1), \\ I_{\text{ion}}(\mathbf{x}, t, v, s_1) &= \frac{1}{\epsilon} \left(v - \frac{v^3}{3} - s_1 \right), \end{aligned} \quad (5)$$

where ϵ is a cell-tissue coupling parameter, γ is a cell recovery rate, and β is a resting potential parameter. The FHN bidomain model parameters used in the experiments reported are $\chi = 1$, $C_m = 1$, $\sigma_i = \sigma_e = 1$, $\epsilon = 0.1$, $\beta = 1$, and $\gamma = 0.5$. The spatial domain is $[0, 70]$, and the simulation takes place on the time interval $[0, 30]$.

3.1 Initiation of Pulse Solutions

To initiate a pulse solution, we initialize the membrane potentials and cell state to their equilibrium values that arise from the parameters specified in the previous sections.

From this state, a simple rectangular stimulus is applied at the left end of the domain to initiate a pulse that will propagate to the right,

$$I_{\text{stim}}(x, t) = \begin{cases} I_{\text{amp}}, & t_0 \leq t \leq t_{\text{stim}}, 0 \leq x \leq x_{\text{stim}}, \\ 0, & \text{otherwise.} \end{cases}$$

The amplitude I_{amp} , duration t_{stim} , and spatial extent x_{stim} of a stimulus that is capable of generating a pulse is dependent on the tissue and cell model parameters. The values used for the numerical experiments are $I_{\text{amp}} = 2$, $t_{\text{stim}} = 2$, and $x_{\text{stim}} = 2$.

3.2 Reference Solution and Error

We compute a reference solution that has converged to D digits of accuracy at N_{ST} points in space and time. The value of D is chosen to be sufficiently large compared to the magnitude of errors observed in the numerical experiments. For our experiment, we have computed a reference solution using the spectral element code Nektar++ [4] and Richardson extrapolation [17] applied in a step-doubling manner in time. The Nektar++ reference solution is computed using the first-order semi-implicit BDF (SBDF1) method with a fixed spatial discretization using $p = 48$ order polynomials on 700 intervals. Richardson extrapolation with 14 levels (i.e., 13 refinements of the time step) is applied starting from $\Delta t = 0.1$. This procedure is performed because Nektar++ does not have any error control facilities.

The error we use is the Mixed Root Mean Square (MRMS) error, which for a quantity W is defined by

$$[\text{MRMS}]_W = \sqrt{\frac{1}{N_{\text{ST}}} \sum_{i=1}^{N_{\text{ST}}} \left(\frac{\hat{W}_i - W_i}{1 + |\hat{W}_i|} \right)^2},$$

where W_i and \hat{W}_i denote the numerical and the reference solutions of component W at space-time point i . The total number of space-time points is taken to be $N_{\text{ST}} = N_x N_t$. As shown in [11], the MRMS error gives a more self-consistent measure of accuracy than the Relative Root Mean Square error when considered across various cell models. The number of equally spaced spatial points for the reference solution is $N_x = 700$ and the number of equally spaced points in the time dimension is $N_t = 21$.

The reference solution calculated in this way has $D = 7$ stable digits and is displayed in Fig. 1. For comparison, the final state of an eBACOLI FHN bidomain solution with tolerances $\text{atol} = \text{rtol} = 10^{-2}$ is given in Fig. 2. Close agreement is indicated by MRMS in Figs. 3 and 4 with the reference solution for an eBACOLI solution using B-spline order $p = 9$ and tolerances $\text{atol} = \text{rtol} = 10^{-9}$. For convenience of comparisons, we use this eBACOLI reference solution henceforth.

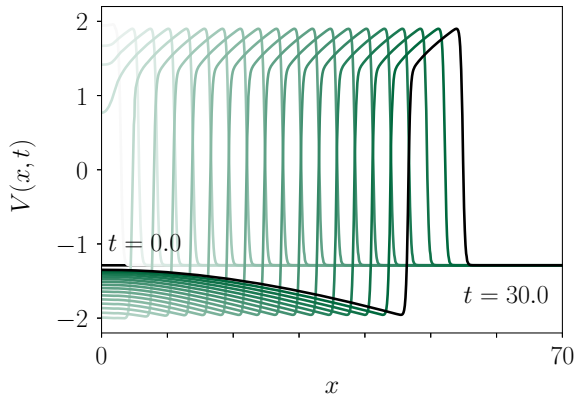


Fig. 1 Snapshots of reference solution for $V(x, t)$ computed using Nektar++ and Richardson extrapolation. Starting from a uniform steady state, the applied stimulus quickly induces a pulse travelling to the right

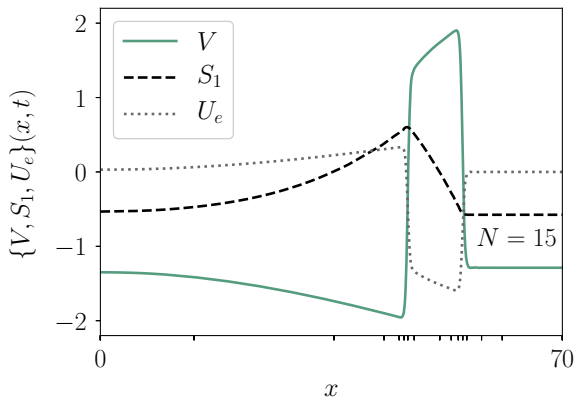


Fig. 2 Snapshot at $t = 30.0$ for all components of an FHN bidomain solution obtained from eBACOLI. Solve parameters are $atol=1e-2$, and polynomial order $p = 5$. Because a loose tolerance is used, this final state requires only 15 mesh subintervals, represented by the x -axis ticks, and corresponds to an $[MRMS]_V$ of 2% when compared to the Nektar++ reference in Fig. 1

3.3 Numerical Solutions

For the purposes of analysis, we compute numerical solutions with eBACOLI and Nektar++ by adjusting solution parameters in prescribed ways. We plot the precision obtained relative to the reference solution as a function of CPU time and tolerance (for eBACOLI) or constant time step Δt (for Nektar++). For eBACOLI, we adjust the $atol$ parameter (and keeping $atol=rtol$). We look at results obtained from three different B-spline expansions with order $p = \{5, 7, 9\}$. For Nektar++, we adjust Δt , keeping both the spatial mesh and the expansion order P constant. We use the

same mesh as that of the reference solution and fix $P = 6$. We look at the first three orders of the SBDF family of time integrators: SBDF1, SBDF2, and SBDF3 [1].

All computations were performed on a computer with an Intel(R) Xeon(R) W3520 @ 2.67GHz CPU, 16GB DDR3 @ 1333MT/s RAM, running 4.15.0-65-generic #74-Ubuntu SMP Tue Sep 17 17:06:04 UTC 2019, and compilers gcc & gfortran (Ubuntu 7.4.0-1ubuntu1 18.04.1) 7.4.0.

4 Results

A number of simulations were performed using the following eBACOLI and Nektar++ methods: three eBACOLI methods using B-spline orders $p = \{5, 7, 9\}$ with various tolerances and three Nektar++ methods using a sixth-order spatial expansion and time integrators SBDF $\{1,2,3\}$ with various constant step sizes. The results of these experiments are given in Figs. 3 and 4 as $[\text{MRMS}]_V$ and $[\text{MRMS}]_{U_c}$ error versus CPU time (Figs. 3a and 4a) and tolerance / Δt (Figs. 3b and 4b).

There are a few observations that can be made from Figs. 3 and 4. First, varying the tolerances in eBACOLI results in less smooth MRMS error and CPU time behavior than varying the timestep Δt in Nektar++. This is expected due to the nature of the adaptive algorithms employed within eBACOLI. The observed non-smooth behaviour can be effectively addressed through the use of strategies based on digital filters [19], but doing so is deemed to be beyond the scope of this study.

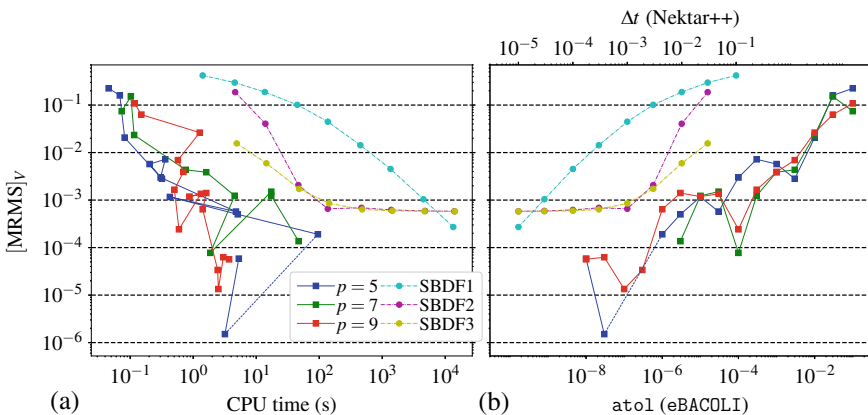


Fig. 3 Precision as a function of CPU time and specified tolerance / time step for three eBACOLI methods with B-spline orders $p = \{5, 7, 9\}$ and three Nektar++ methods with sixth-order spatial expansions and time integrators SBDF $\{1,2,3\}$. **a** Precision-work diagram for a given $[\text{MRMS}]_V$ error. **b** Precision for different tolerance values (for eBACOLI) and time steps (for Nektar++). The bottom axis indicates $\text{atol}=\text{rtol}$ for eBACOLI solutions, and the top axis indicates Δt for Nektar++ solutions

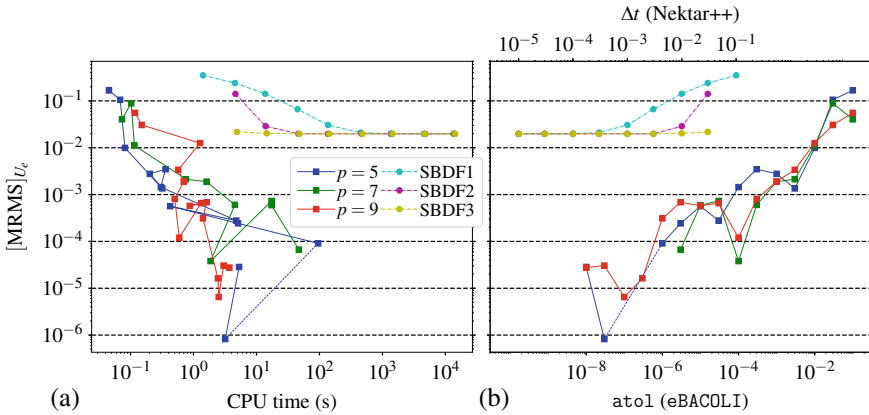


Fig. 4 Precision as a function of CPU time and specified tolerance / time step for three eBACOLI methods with B-spline orders $p = \{5, 7, 9\}$ and three Nektar++ methods with sixth-order spatial expansions and time integrators SBDf{1,2,3}. **a** Precision-work diagram for a given $[MRMS]_{U_e}$ error. **b** Precision for different tolerance values (for eBACOLI) and time steps (for Nektar++). The bottom axis indicates $atol=rtol$ for eBACOLI solutions, and the top axis indicates Δt for Nektar++ solutions

Second, for the Nektar++ simulations, $[MRMS]_{U_e}$ is generally larger than $[MRMS]_V$, and convergence towards the reference solution stagnates. For the eBACOLI simulations, $[MRMS]_{U_e}$ is generally slightly smaller than $[MRMS]_V$, and convergence towards the reference solution does not stagnate. Limiting the order of the spatial expansion in Nektar++ for these simulations to $P = 6$ is likely an important determining factor in their stagnation. Because eBACOLI simulations are based on satisfying given tolerances, fixing the order of the spatial expansion does not result in this behavior.

Finally, the eBACOLI solutions are generally obtained in at least an order of magnitude less time than the Nektar++ solutions at a given MRMS error level, with differences reaching as high as four orders of magnitude. There is also no clear winner among the different eBACOLI solutions in terms of B-spline order. The highest-order eBACOLI simulation can achieve results with tolerances as low as 10^{-9} (the reference solution), but the true accuracy of such a solution cannot be determined by the methods employed in this study.

5 Conclusions

Adaptive B-spline collocation is shown to be an effective method for one-dimensional systems of time-dependent PDEs that consist of parabolic and elliptic PDEs coupled with ODEs as described by the bidomain model Eq. (1). Besides providing automatic error control, the adaptive B-spline solutions obtained using eBACOLI are obtained

much more efficiently, up to four orders of magnitude faster for a given MRMS error level, than spectral element solutions in Nektar++ with non-adaptive spatial and time discretizations. High-accuracy solutions to cardiac bidomain problems can be obtained so quickly by eBACOLI that we recommend its use for efficient reference solution generation for one-dimensional cardiac bidomain problems.

Future directions of this work include (i) increasing the performance of eBACOLI by allowing faster construction of Jacobians and solves of the resulting almost block diagonal (ABD) linear systems [16] by replacing COLROW [5] with a parallel ABD solver like RSCALE [14] and (ii) building a general one-dimensional cardiac solver based on eBACOLI that interfaces with all cell models from the CellML library [2].

References

1. Ascher, U.M., Ruuth, S.J., Wetton, B.T.R.: Implicit-explicit methods for time-dependent partial differential equations. *SIAM J. Num. Anal.* **32**(3), 797–823 (1995)
2. Auckland Bioengineering Institute: The CellML project (2011). <http://www.cellml.org/>
3. de Boor, C.: *A Practical Guide to Splines*, vol. 27. Springer, New York, USA (1978)
4. Cantwell, C.D., Moxey, D., Comerford, A., Bolis, A., Rocco, G., Mengaldo, G., De Grazia, D., Yakovlev, S., Lombard, J.E., Ekelschot, D., Jordi, B., Xu, H., Mohamied, Y., Eskilsson, C., Nelson, B., Vos, P., Biotto, C., Kirby, R.M., Sherwin, S.J.: Nektar plus plus : an open-source spectral/hp element framework. *Comput. Phys. Commun.* **192**, 205–219 (2015). <https://doi.org/10.1016/j.cpc.2015.02.008>
5. Diaz, J.C., Fairweather, G., Keast, P.: Fortran packages for solving certain almost block diagonal linear systems by modified alternate row and column elimination. *ACM Trans. Math. Softw.* **9**(3), 358–375 (1983). <https://doi.org/10.1145/356044.356053>
6. Ethier, M., Bourgault, Y.: Semi-implicit time-discretization schemes for the bidomain model. *SIAM J. Numer. Anal.* **46**(5), 2443–2468 (2008). <https://doi.org/10.1137/070680503>
7. FitzHugh, R.: Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1**(6), 445–466 (1961)
8. Green, K.R., Spiteri, R.J.: Extended BACOLI: solving one-dimensional multiscale parabolic pde systems with error control. *ACM Trans. Math. Softw.* **45**(1), 8:1–8:19 (2019). <https://doi.org/10.1145/3301320>
9. Hooke, N., Henriquez, C., Lanzkron, P., Rose, D.: Linear algebraic transformations of the bidomain equations: implications for numerical methods. *Math. Biosci.* **120**(2), 127–145 (1994). [https://doi.org/10.1016/0025-5564\(94\)90049-3](https://doi.org/10.1016/0025-5564(94)90049-3). <http://www.sciencedirect.com/science/article/pii/0025556494900493>
10. Keener, J.P., Bogar, K.: A numerical method for the solution of the bidomain equations in cardiac tissue. *Chaos Interdisc. J. Nonlinear Sci.* **8**(1), 234–241 (1998). <https://doi.org/10.1063/1.166300>
11. Marsh, M.E., Ziaratgahi, S.T., Spiteri, R.J.: The secrets to the success of the Rush-Larsen method and its generalizations. *IEEE Trans. Biomed. Eng.* **59**(9), 2506–2515 (2012). <https://doi.org/10.1109/TBME.2012.2205575>
12. Nagumo, J., Arimoto, S., Yoshizawa, S.: An active pulse transmission line simulating nerve axon. *Proc. IRE* **50**(10), 2061–2070 (1962). <https://doi.org/10.1109/JRPROC.1962.288235>
13. Niederer, S.A., et al.: Verification of cardiac tissue electrophysiology simulators using an N-version benchmark. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **369**(1954), 4331–4351 (2011). <https://doi.org/10.1098/rsta.2011.0139>
14. Pancer, R.N.: *The Parallel Solution of ABD Systems Arising in Numerical Methods for BVPs for ODEs*. Ph.D. thesis, University of Toronto (2006)

15. Petzold, L.R.: A description of DASSL: a differential-algebraic system solver. In: Scientific computing (Montreal, Que., 1982), IMACS Trans Sci Comput., I, pp. 65–68. IMACS, New Brunswick, NJ (1983)
16. Pew, J., Li, Z., Muir, P.: Algorithm 962: BACOLI: B-spline adaptive collocation software for PDEs with interpolation-based spatial error control. *ACM Trans. Math. Softw.* **42**(3), 25 (2016). <https://doi.org/10.1145/2818312>
17. Richardson, L.F.: On the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences An Interdisciplinary Journal, vol. 83(563), pp. 335–336 (1910). <https://doi.org/10.1098/rspa.1910.0020>. <http://rspa.royalsocietypublishing.org/content/83/563/335>
18. Roth, B.J.: Action potential propagation in a thick strand of cardiac muscle. *Circ. Res.* **68**(1), 162–173 (1991)
19. Söderlind, G., Wang, L.: Adaptive time-stepping and computational stability. *J. Comput. Appl. Math.* **185**(2), 225–243 (2006). <https://doi.org/10.1016/j.cam.2005.03.008>. <https://doi.org/cyber.usask.ca/10.1016/j.cam.2005.03.008>
20. Sundnes, J., Lines, G.T., Cai, X., Nielsen, B.F., Mardal, K.A., Tveito, A.: Computing the Electrical Activity in the Heart. Springer, Berlin (2006)
21. Tung, L.: A bi-domain model for describing ischemic myocardial dc potentials. Ph.D. thesis, Massachusetts Institute of Technology (1978)

A Computational Comparison of Three Methods for Solving a 1D Boundary Value Inverse Problem



Kimberly M. Levere, Bryson Boreland, and John Dewhurst

Abstract A goal of many inverse problem techniques is to find unknown parameter values $\lambda \in \Lambda$ that produce a solution to the forward problem, u_λ , that lies “close” to a known solution, u . Mathematically speaking, these techniques wish to minimize the approximation error subject to these parameters,

$$\min_{\lambda \in \Lambda} \|u - u_\lambda\|.$$

A number of different inverse problem techniques have been developed for solving such a problem. In this paper we briefly discuss three methods for solving inverse problems in the ODE setting: Tikhonov Regularization, Landweber-Fridman iteration, and the more recent Collage-Coding method. We compare and contrast the methods by applying each of them to the same example. The accuracy, robustness, and efficiency of each method is then explored.

Keywords Inverse problem · Optimization · Parameter estimation · Collage theorem · Regularization

1 Introduction

A number of different inverse problem techniques have been developed for solving inverse problems with a variety of different applications. In this paper we focus our efforts on inverse problems for ODEs. Perhaps the most well-known inverse prob-

K. M. Levere (✉) · B. Boreland · J. Dewhurst
University of Guelph, 50 Stone Road E., Guelph, ON N1G 2W1, Canada
e-mail: klevere@uoguelph.ca

B. Boreland
e-mail: bborelan@uoguelph.ca

J. Dewhurst
e-mail: dewhurstj@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_29

lem methods are regularization techniques, in particular, Tikhonov Regularization [11]. Regularization techniques often cast the above problem in terms of a system $A\kappa = f$, where A is an operator containing known information, f contains known information, and κ contains the unknown parameters. As the operator A is often not invertible, or does not have a bounded inverse, the problem of solving for κ is ill-posed. Regularization schemes replace the ill-posed operator with a “nearby” well-posed operator and then add a penalty (or regularization) term to correct for this adjustment.

Iterative Schemes are also prevalent in the inverse problem literature. As their name implies, an iteration scheme is developed that is designed to converge to a solution (after a large number of iterations) that is “close” to the true parameter values. One such iteration scheme is explored in this paper, Landweber-Fridman iteration [2].

A more recent inverse problem technique is that of Collage-Coding [8]. The idea behind this technique is to bound the approximation error above by another distance that is more easily minimized. When working with ODEs, this upper bound is achieved using the Collage Theorem, a consequence of the well-known Banach’s Fixed Point Theorem,

$$\|u - u_\lambda\| \leq \frac{1}{1 - c_\lambda} \|u - T_\lambda u\|, \quad (1)$$

where T_λ is a c_λ -contractive, space preserving operator that depends on the unknown parameters λ . By minimizing the so-called collage distance $\|u - T_\lambda u\|$ (ensuring that c_λ is bounded away from 1), one can ensure that the approximation error is indeed small. Relying instead on the Lax-Milgram Representation Theorem, Collage-Coding methods have been developed for both linear and nonlinear PDEs.

In Sects. 2–4, we explore each of the above mentioned methods in turn, outlining their basic approach. Each method is then used to solve the example problem

$$\begin{aligned} \frac{d}{dx} \left(\kappa(x) \frac{du}{dx} \right) &= -f(x) \text{ in } \Omega = (0, 1), \\ u &= 0, \text{ on } \partial\Omega \end{aligned} \quad (2)$$

to highlight details and numerical considerations of each approach. Finally, in Sect. 5 we compare and contrast the methods and discuss their strengths and weaknesses.

2 Regularization Methods

Regularization schemes attempt to handle the challenge of minimizing the approximation error head-on. The problem is first recast as an operator equation $A\kappa = f$, where A and f contain only known information, and κ contains the unknown parameter values. In this case then, the approximation error becomes the distance between $A\kappa$ and f in an appropriate norm, $\|A\kappa - f\|$. If the operator A is invertible so that

$\kappa = A^{-1}f$, and this κ depends continuously on f then the problem is said to be well-posed. If either of these conditions fails, the problem is called ill-posed. Depending on the severity of the ill-posedness, solutions can be found in a variety of ways. For instance, if κ does not depend continuously on f , one can orthogonally project f onto the closure of the range of A , $\Pi f \in R(\bar{A})$. We then can achieve a so-called generalized solution, which is defined to be the κ that, under the operator A , is closest in norm to Πf . Such solutions can be shown to exist and be unique, see [3] for further details.

2.1 Tikhonov Regularization

If the problem suffers not only from a lack of continuous dependence but the operator A is also not invertible, then the problem is called genuinely ill-posed. In this case, regularization methods can be of use, in particular, we will discuss Tikhonov Regularization. The idea is to replace the ill-posed operator by a nearby well-conditioned operator. To arrive at this operator, we recall that we wish to minimize the approximation error

$$\begin{aligned} \min_{\kappa} \frac{1}{2} \|A\kappa - f\|^2 &= \min_{\kappa} \frac{1}{2} (A\kappa - f)^*(A\kappa - f) \\ &= \min_{\kappa} \frac{1}{2} (\kappa^* A^* A \kappa - 2\kappa^* A^* f + f^* f), \end{aligned}$$

where A^* denotes the adjoint operator of A . Calculus tells us that the values of κ that accomplish this task satisfy

$$\begin{aligned} \frac{\partial}{\partial \kappa_i} (\kappa^* A^* A \kappa - 2\kappa^* A^* f + f^* f) &= 0 \\ \implies A^* A \kappa &= A^* f. \end{aligned} \tag{3}$$

Unfortunately, the operator $A^* A$ is often not invertible. To remedy this, we replace this operator by the perturbed operator $A^* A + \alpha I$, where $\alpha > 0$ (called the regularization parameter) and I denotes the identity operator. By choosing $\alpha > 0$ such that all spectral values are positive the perturbed operator is indeed invertible. By substituting the perturbed operator into (3) we arrive at the Tikhonov approximation to the generalized solution

$$\kappa = (A^* A + \alpha I)^{-1} A^* f.$$

One can prove that under particular conditions that as $\alpha \rightarrow 0^+$ the Tikhonov approximation indeed converges to the generalized solution. See, for instance, [2] and [11] for further details.

2.2 An Example Problem

We apply the technique of Tikhonov Regularization to the BVP given in (2). To recast this continuous problem in terms of a discrete operator A , r points are chosen along the domain $\Omega = [0, 1]$ (in our case, uniformly although this is not necessary). Derivatives are approximated by finite difference approximations and thus A is a matrix containing combinations of discrete data values, u_i , $0 \leq i \leq r$, while f is a vector containing discrete values of the function $f(x)$, f_i and κ contains unknown values of the function $\kappa(x)$ at each of the r discrete locations on Ω , κ_i . In order to verify our results, we assume $f(x) = 96x^3 + 12x^2 + 48x - 8$ and $\kappa_{\text{True}}(x) = 3x^3 + 2x^2 + 4x + 1$ which imposes $u_{\text{True}}(x) = 4x(1 - x)$. To construct faux data, we then sample $u_{\text{True}}(x)$ at the r discrete locations along Ω , possibly adding low amplitude Gaussian noise, ε to simulate experimental error. For the inverse problem we imagine that we only have access to the data values u_i , and the discrete values f_i . We seek discrete values κ_i that minimize the Tikhonov approximation error $\|(A^*A + \alpha I)\kappa - A^*f\|$. Subsequently, a polynomial is fit to the κ_i via least squares to form the continuous approximation $\kappa_{\text{Tikhonov}}(x)$. Using this continuous approximation, the original BVP is solved forward to find the Tikhonov approximation of u , $u_{\text{Tikhonov}}(x)$. Numerous trials were run testing the impact of using different degrees for $\kappa_{\text{Tikhonov}}(x)$. The method seemed to correctly identified the appropriate degree making higher degree coefficients negligible. Note that the regularization parameter was chosen using Mozorov’s principle which is discussed in detail in [2]. The results of several trials are summarized in Table 1. In Fig. 1 both the recovered and true values of $\kappa(x)$ are plotted for comparative purposes.

As one might expect, increasing the number of sample points, r , produces better results with respect to the approximation error of $\kappa(x)$. For a fixed number of sample points, the approximation error of $u(x)$ shows strong tolerance to noise, as it remains fairly consistent across different amplitudes of noise.

Table 1 Results for Tikhonov Regularization applied to (2) for various values of r and ε

r	ε	α	$\ \kappa_{\text{Tikhonov}} - \kappa_{\text{True}}\ _{L^2(\Omega)}$	$\ u_{\text{Tikhonov}} - u_{\text{True}}\ _{L^2(\Omega)}$
5	0	0.30951×10^{-9}	0.99591	0.19756
	0.01	0.17011×10^{-3}	0.99548	0.19705
	0.1	0.17072×10^{-2}	0.98979	0.19205
10	0	5.94456×10^{-10}	0.65853	0.28319
	0.01	0.12131×10^{-4}	0.65772	0.28254
	0.1	0.12166×10^{-3}	0.65165	0.27716
15	0	0.42286×10^{-14}	0.59133	0.27981
	0.01	0.25567×10^{-14}	0.58914	0.27672
	0.1	0.39503×10^{-14}	0.57156	0.26784

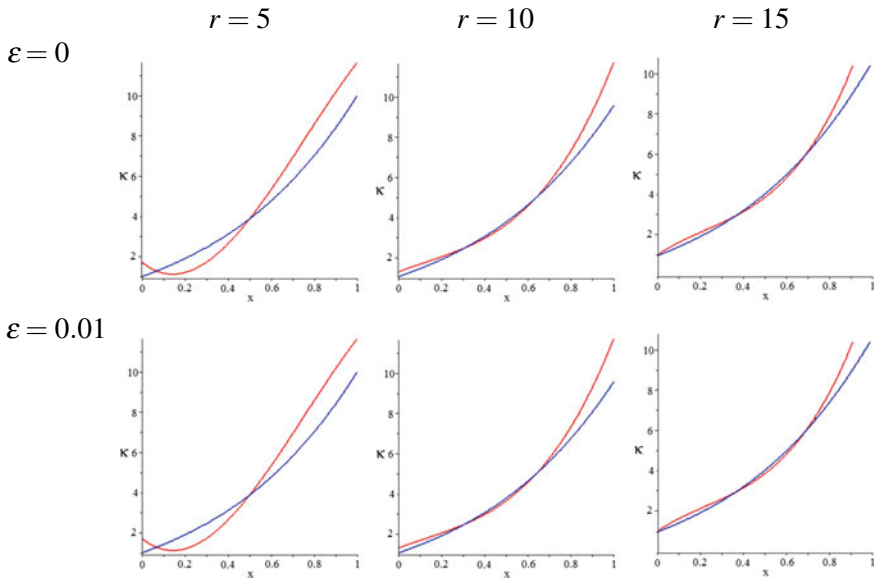


Fig. 1 Plots of $\kappa_{\text{Tikhonov}}(x)$ (red) and $\kappa_{\text{True}}(x)$ (blue) for various values of r and ε

3 Iteration Schemes

Much like regularization methods, iteration schemes work with discrete values. Using an initial guess or seed value, all future values can be found based on previous values. The sequence of values that are generated can be shown to converge (under certain conditions) to a solution (generalized in our case) of the original problem.

3.1 Landweber-Fridman Iteration

Utilizing the same setup as regularization methods, iterative schemes recast continuous problems in terms of the operator equation $A\kappa = f$, where A and f contain only known information, and κ contains the unknown parameter values. As before, to minimize the approximation error, $\|A\kappa - f\|$ one must find κ such that $A^*A\kappa = A^*f$. Working with this requirement, first rearrange and then multiply by a parameter $\beta > 0$ (which will serve as the stepsize for the iterative scheme)

$$0 = \beta(A^*f - A^*A\kappa).$$

Finally, adding κ to both sides and appropriately subscripting suggests the iterative scheme

$$\kappa_{n+1} = \kappa_n + \beta(A^*f_n - A^*A\kappa_n),$$

where $n = 0, \dots$ and κ_0 is the seed value. With an appropriate choice for β one can show that the sequence $\{\kappa_n\}_{n=0}^\infty$ converges to the generalized solution, see [2] for further details.

3.2 Revisited: An Example Problem

We consider again the BVP problem (2) presented in the introduction.

To construct the iteration scheme, we use the same process to discretize this continuous problem. Motivated by convergence requirements, we choose $\beta = \frac{1}{\|A\|} > 0$ as our stepsize for the method and a stopping criteria of $\|A\kappa - f\| \leq 10^{-10}$. At the end of iteration, a polynomial was fit to the κ found by iteration, $\kappa_{\text{Landweber}}(x)$. Using this continuous approximation, we then solved (2) to determine the approximate solution $u_{\text{Landweber}}(x)$. As was the case with Tikhonov Regularization, several degrees for $\kappa_{\text{Landweber}}(x)$ were tested. When using higher degree polynomials than necessary, these higher order coefficients were found to be negligible. The results for various values of r and ε are found in Table 2. In Fig. 2 both the recovered and true values of $\kappa(x)$ are plotted for comparative purposes.

Landweber-Fridman iteration produces very similar quality results to that of Tikhonov Regularization. As expected, increasing r produces better results for our recovered κ while the introduction of noise hinders this method although it is rather robust to low amplitude noise.

Table 2 Results for Landweber-Fridman iteration applied to (2) for various values of r and ε

r	ε	$\ \kappa_{\text{Landweber}} - \kappa_{\text{True}}\ _{L^2(\Omega)}$	$\ u_{\text{Landweber}} - u_{\text{True}}\ _{L^2(\Omega)}$
5	0	0.99591	0.19756
	0.01	0.99655	0.19730
	0.1	0.99917	0.19402
10	0	0.65853	0.28319
	0.01	0.65812	0.28264
	0.1	0.65663	0.27859
15	0	0.59133	0.27981
	0.01	0.59103	0.27940
	0.1	0.58968	0.27614
20	0	0.53069	0.26455
	0.01	0.53031	0.26404
	0.1	0.52837	0.26037

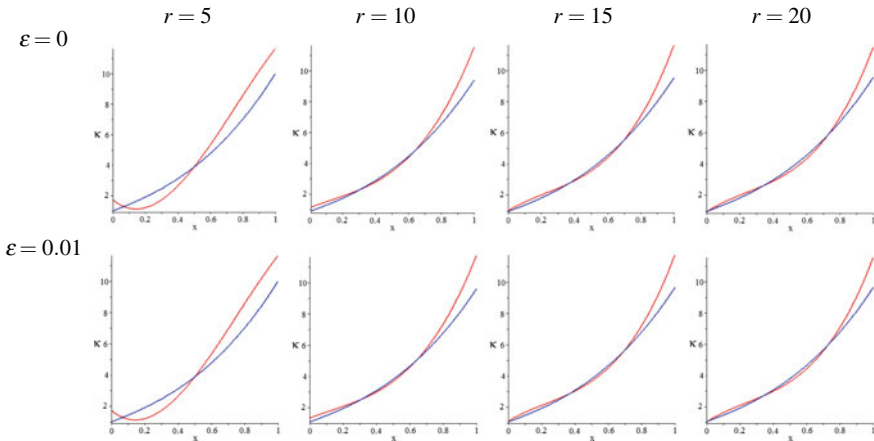


Fig. 2 Plots of $\kappa_{\text{Landweber}}(x)$ (red) and $\kappa_{\text{True}}(x)$ (blue) for various values of r and ε

4 Collage-Coding Methods

The Collage-Coding technique was first proposed in 1999 and applied to inverse problems for ODEs in [8]. Unlike regularization methods or iterative schemes that tackle the approximation error head-on, the Collage-Coding method instead bounds the approximation error above by the so-called collage distance which is, in practice, easier to minimize. By minimizing the collage distance (subject to a few conditions), the approximation error will be controlled. The construction of the collage distance relies on the same hypotheses required for establishing existence and uniqueness of solutions to ODEs. For completeness, we state Banach’s Fixed Point Theorem which is commonplace in this regard.

Theorem 1 (Banach’s Fixed Point Theorem) *Let $(X, \| \cdot \|_X)$ be a Banach space and let $T : X \rightarrow X$ be a contractive operator with contraction factor $c \in [0, 1)$. Then there exists a unique fixed point $\bar{u} \in X$ such that $T\bar{u} = \bar{u}$. Moreover, for any $u \in X$, $\|T^{os}u - \bar{u}\|_X \rightarrow 0$ as $s \rightarrow \infty$.*

Proof of this theorem can be found in [12], for instance. A simple consequence of this theorem, the Collage Theorem, will establish the aforementioned upper bound on the approximation error.

Theorem 2 (Collage Theorem) *Let $(X, \| \cdot \|_X)$ be a Banach space and $T : X \rightarrow X$ be a contractive operator with contraction factor $c \in [0, 1)$ and unique fixed point $\bar{u} \in X$. Then*

$$\|u - \bar{u}\|_X \leq \frac{1}{1 - c} \|u - Tu\|_X.$$

The proof of this theorem can be found in [1]. By minimizing the collage distance, $\|u - Tu\|_X$, provided c is bounded away from 1, guarantees that the approximation

error is indeed controlled. This is the theme of a wide variety of Collage-Coding methods that have since been developed for treating a variety of different problems, see for instance [4]–[10].

4.1 Revisited: An Example Problem

Once again, we apply the current method to the example problem (2) presented in the introduction.

We will work on the complete metric space $(C^1(\bar{I}), \|\cdot\|_\infty)$. Since the infinity norm is computationally cumbersome, we utilize the fact that $C^1(\bar{\Omega}) \subset \mathcal{L}^2(\bar{\Omega})$ and work instead with the \mathcal{L}^2 norm. As before, we assume $f(x) = 96x^3 + 12x^2 + 48x - 8$ and that we have been given observational data for $u(x)$ (in this case generated by sampling the function $u(x) = 4x(1-x)$ at r discrete locations along the domain $[0, 1]$). A second-degree polynomial target function, $u_{\text{target}}(x)$ is then fit to this data using a least squares procedure. With this setup, $\kappa_{\text{true}}(x) = 3x^3 + 2x^2 + 4x + 1$, although in practice this would not be known so we will only use this to check the accuracy of our method. A suitable choice for a contractive, space-preserving operator T for this problem is a Picard operator. For our BVP, this is found by first applying the product rule on the left-hand side of the ODE, integrating twice is given by

$$Tu(x) = u_{\text{target}}(0) + u'_{\text{target}}(0)x - \int_0^x (x-s) \left(\frac{f(s) + \kappa'_{\text{collage}}(s)u'_{\text{target}}(s)}{\kappa_{\text{collage}}(s)} \right) ds$$

where we assume the form $\kappa_{\text{collage}}(x) = \lambda_3x^3 + \lambda_2x^2 + \lambda_1x + \lambda_0$ and seek the values of λ_i that minimize the collage distance. Different degrees for $\kappa_{\text{collage}}(x)$ were tested, with higher order coefficients recovered as negligible values. Certainly, one could instead express $\kappa_{\text{collage}}(x)$ in terms of an appropriate basis to avoid the use of polynomials and the choice of their degree altogether. Note that one can prove that the operator is indeed contractive and space-preserving (see [7]). As the parameters of the problem appear in a complex way in our operator T (and thus, also in our collage distance), we apply the Taylor expansion to both of the terms $\frac{\kappa'(x)}{\kappa(x)}$

and $\frac{f(x)}{\kappa(x)}$ to simplify integration. Depending on the complexity of the operator a variety of minimization schemes may be employed. In this case a gradient descent was used to find a solution. The computational results of the Collage-Coding method are shown in Table 3 with graphical results in Fig. 3.

Perhaps not surprisingly, as we are given more data (larger r) our results improve. Likewise, we see a decrease in the accuracy of our results as we increase the amplitude of noise, ε .

Table 3 Results for Collage-Coding applied to (2) for various values of r and ε with the collage distance given by $F(\lambda)$

r	ε	$\ \kappa_{collage} - \kappa_{true}\ _{L^2(\Omega)}$	$\ u_{collage} - u_{true}\ _{L^2(\Omega)}$	$F(\lambda)$
5	0	0.40638×10^{-2}	0.10628×10^{-2}	0.13146×10^{-2}
	0.01	0.15850×10^{-1}	0.13428×10^{-2}	0.13644×10^{-2}
	0.1	0.14369	0.82542×10^{-2}	0.13771×10^{-1}
10	0	0.22166×10^{-2}	0.41095×10^{-3}	0.72132×10^{-3}
	0.01	0.29545×10^{-2}	0.52345×10^{-3}	0.75067×10^{-3}
	0.1	0.17485×10^{-1}	0.31956×10^{-2}	0.75115×10^{-2}
15	0	0.15239×10^{-2}	0.23054×10^{-3}	0.46148×10^{-3}
	0.01	0.27604×10^{-2}	0.28666×10^{-3}	0.50843×10^{-3}
	0.1	0.87824×10^{-2}	0.17322×10^{-2}	0.51642×10^{-2}
20	0	0.11611×10^{-2}	0.15201×10^{-3}	0.36302×10^{-3}
	0.01	0.18407×10^{-2}	0.18431×10^{-3}	0.39011×10^{-3}
	0.1	0.84030×10^{-2}	0.10645×10^{-2}	0.39346×10^{-3}

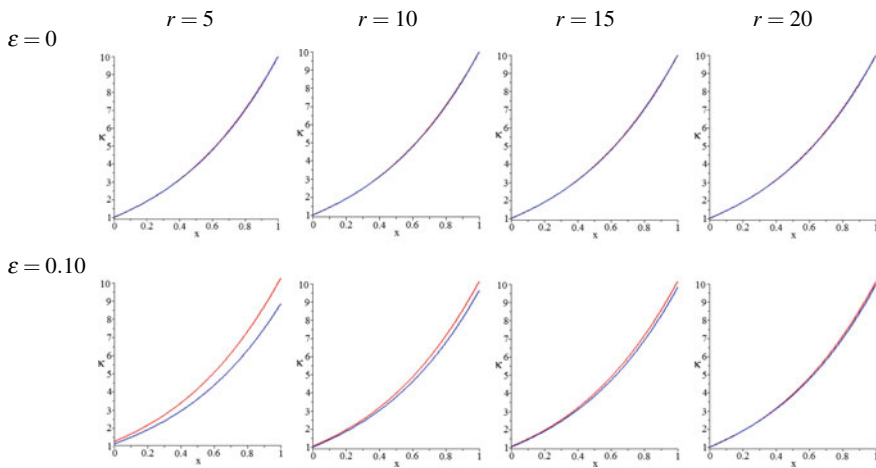


Fig. 3 Plots of $\kappa_{collage}(x)$ (red) and $\kappa_{True}(x)$ (blue) for various values of r and ε

5 Conclusions and Future Work

While each of the methods discussed in this paper provide viable options for finding a solution, certainly there are advantages to each. Tikhonov Regularization and Landweber-Fridman iteration have very simplistic constructions with easy-to-check conditions for the parameters α and β . In the case of the Collage-Coding method, the contraction factor c is difficult to check a priori since in general it will depend on the unknown function $\kappa(x)$. For simplicity, in this paper we chose to express $\kappa(x)$ as

a polynomial, which, in spirit acts as a regularizer on its own. Certainly it is possible to work in terms of an appropriate basis to avoid choosing any particular form for $\kappa(x)$. Bases such as the “hat basis” or the basis of “hexagonal based pyramids” have been employed in other Collage-Coding work. Such choices have not been found to starkly impact the quality of results. Likewise, the target function $u_{\text{target}}(x)$ can also be expressed in terms of a basis so that the use of an arbitrary degree polynomial is not necessary.

Computationally, Tikhonov Regularization was by far the most expensive of the methods investigated. It was unable to produce results for the $r = 20$ case without resorting to higher computing power. Landweber-Fridman iteration was slightly less computationally expensive but required many iterations to reach a solution that indeed converged to the generalized solution. Collage-Coding was the least expensive computationally, finding its results in less than half the time of the iteration scheme.

All three methods were robust to noisy data provided that the number of data points, r , was large. In terms of accuracy, for this example the results of Tikhonov Regularization and Landweber-Fridman iteration were relatively similar in terms of approximation error. Collage-Coding produced much smaller approximation errors than that of the other two methods. Certainly, we cannot conclude that this will be the case more globally when used on other more complex problems, but it is worth noting that gains in accuracy can be seen with this method. Further investigation should be conducted on a variety of problems to more completely investigate the pros and cons of each method.

References

1. Barnsley, M.F., Ervin, V., Hardin, D., Lancaster, J.J.: Solution of an inverse problem for fractals and other sets. *Proc. Natl. Acad. Sci.* **83**, 1975–1977 (1985)
2. Groetsch, C.W.: *Inverse Problems in the Mathematical Sciences*. Vieweg, Wiesbaden (1993)
3. Ivanov, V.K., Vasin, V.V., Tanana, V.P.: *Theory of Ill-Posed Problems and its Applications*. Moscow (1978)
4. Kunze, H., La Torre, D., Levere, K.M., Ruiz Galan, M.: Inverse problems via the generalized collage theorem for vector-valued Lax-Milgram-based variational problems. *Math. Prob. Eng.* **2015** 2015
5. Kunze, H.E., La Torre, D., Levere, K.M., Vrscay, E.R.: Solving inverse problems for deterministic and random delay integral equations using the collage method. *Int. J. Math. Stat.* **11**(1), 1–11 (2012)
6. Kunze, H., La Torre, D., Vrscay, E.R.: A generalized collage method based upon the Lax-Milgram functional for solving boundary value inverse problems. *Nonlinear Anal. Theory Methods Appl.* **71**(12), 1337–1343 (2009)
7. Kunze, H., Murdock, S.: Solving inverse two-point boundary value problems using collage coding. *Inverse Prob.* **22**, 1179–1190 (2006)
8. Kunze, H., Vrscay, E.R.: Solving inverse problems for ordinary differential equations using the Picard contraction mapping. *Inverse Prob.* **15**, 745–770 (1999)
9. Levere, K.M., Kunze, H., La Torre, D.: A collage-based approach to solving inverse problems for second-order nonlinear hyperbolic PDEs. *Commun. Nonlinear Sci. Numer. Simul.* **29**, 283–299 (2015)

10. Levere, K.M., Van De Walker, B.: Solving inverse problems for fractional ODEs via the collage theorem. *Recent Adv. Math. Stat. Methods (AMMCS 2017 Conference Proceedings)* **259**, 127–136 (2018)
11. Tikhonov, A., Arsenin, V.: *Solutions of Ill-posed Problems*. V.H. Winston & Sons, Washington (1977)
12. Zeidler, E.: *Applied Functional Analysis: Applications to Mathematical Physics*. Springer-Verlag, New York (1995)

A Comparison of Turbulence Generated by 3DS Sparse Grids with Different Blockage Ratios and Different Co-frame Arrangements



M. Syed Usama and Nadeem A. Malik

Abstract A new type of grid turbulence generator, the 3D sparse grid (3DS), is a co-planar arrangement of co-frames each containing a different length scale of grid elements [Malik, N. A. US Patent No. US 9,599,269 B2 (2017)] and possessing a much bigger parameter space than the flat 2D fractal square grid (2DF). Using DNS we compare the characteristics of the turbulence (mean flow, turbulence intensity, energy spectrum) generated by different types of 3DS grids. The peak intensities generated by 3DS can exceed the peaks generated by the 2DF by 80%; we observe that a 3DS with blockage ratio 24% produces turbulence similar to the 2DF with blockage ratio 32% implying lower energy input for the same turbulence.

Keywords Turbulence · Sparse grid · Turbulence generator · Fractal · Multi-scale · Blockage ratio · Mixing

1 Introduction

The generation and control of turbulence is one of the most important challenges in fluid mechanics, with applications ranging from drag reduction to mixing in chemical reactors. A promising innovation in recent times has been the design of new types of turbulence generating grids which are different to the classical regular grid (RG), Fig. 1a. The RG grids have bars of fixed thickness and flow passages of fixed size. A new grid type is a multi-scale arrangement of bars of varying thicknesses that produce flow passages of various sizes. Typically, the bar thicknesses and flow passages are in a self-similar configuration in a two-dimensional plane, such as the 2D square fractal grid (2DF), Fig. 1b. A key feature of multi-scale 2D grids is that they produce multiple scales of turbulence at once in the grid plane, which alters the turbulence generated

M. S. Usama
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

N. A. Malik (✉)
Department of Mechanical Engineering, Texas Tech University, Lubbock, TX 74909, USA
e-mail: nadeem.malik@ttu.edu; nadeem_malik@cantab.net

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_30

325

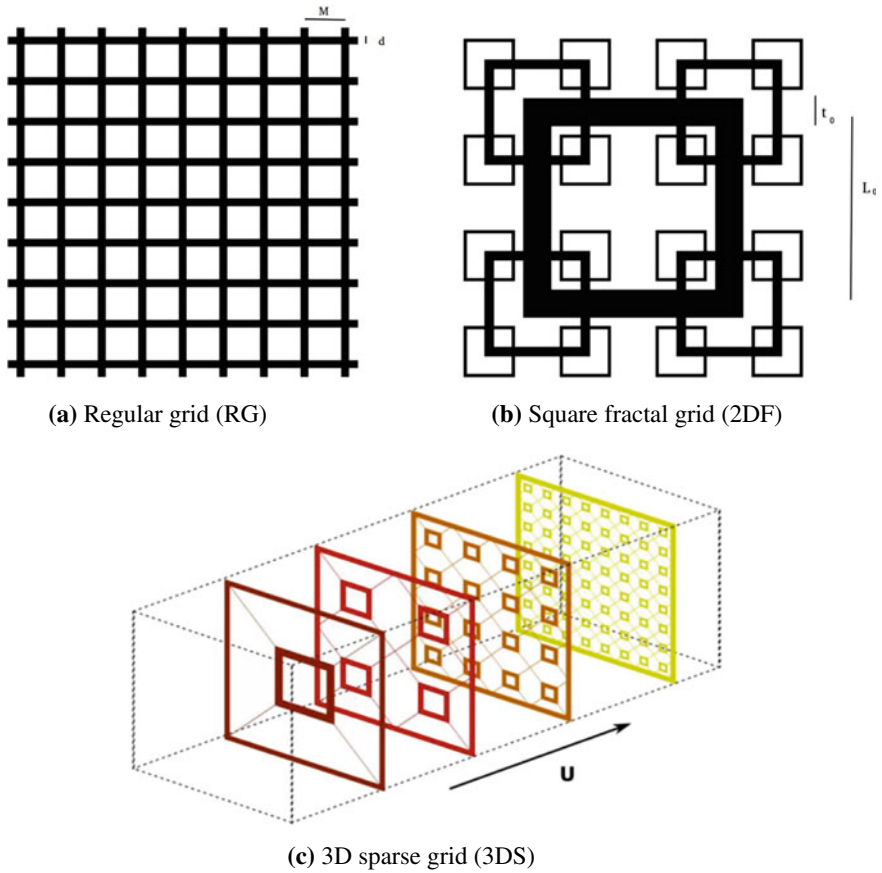


Fig. 1 Different types of grid

[1–5] compared to RG; in particular the peak turbulence intensity is enhanced for the same blockage ratio [3].

Consider a rectangular flow channel or conduit with a turbulence grid placed close to the entrance. A defining characteristic of turbulence generating grids is the blockage ratio (i.e. the solidity), σ , which is the surface area of all the bar elements, S_{elts} divided by the planar cross-sectional area A of the channel,

$$\sigma = \frac{S_{elts}}{A} \tag{1}$$

In the RG and 2DF, σ is a single value; in [3] a three-generation 2DF was presented with $\sigma_{2DF} = 0.32$ (or 32%). σ is important for flow passage; for the same volumetric flow rate you need higher pressure gradient in a channel with higher solidity, which means more energy input requirement. Thus, an important goal in mixing is to opti-

mize the balance between energy input (or $\partial P/\partial x$), the solidity σ , and turbulence generation.

A recent innovation in grid generated turbulence, the Sparse 3D Multi-Scale Grid Turbulence Generator, or 3D sparse grid (3DS) for short [6, 7], has excited interest in the turbulence community because of its potential to alter and control turbulence characteristics even more than the 2DF. The 3DS separates each generation of length scale of grid elements into its own co-frame in overall co-planar arrangement, Fig. 1c, which produces a 3D ‘sparse’ grid system. Each generation of grid elements produces a turbulent wake pattern that interacts with the other wake patterns downstream. The length scale of the grid elements from co-frame to co-frame can be in any geometric ratio, although a fractal pattern across the generations is a popular choice. The spacing between successive co-frames r_1, r_2, \dots are new parameters which do not exist in a non-sparse single frame 2D grid system. If each co-frame is located at $[x_0, x_1, x_2, \dots]$, then $r_1 = x_1 - x_0, r_2 = x_2 - x_1$, etc. Each co-frame has a blockage ratio, $\sigma_0, \sigma_1, \sigma_2, \dots$. We define the overall (or maximum) blockage ratio of the 3DS system, σ_{3DS} , to be the maximum of this set of values,

$$\sigma_{3DS} = \text{Max}\{\sigma_0, \sigma_1, \sigma_2, \dots\}. \quad (2)$$

Thus, for the same value of σ_{3DS} there are an infinite number of possible 3DS configurations since the σ'_i s can take continuous values, provided $0 < \sigma_i \leq \sigma_{3DS}$; at least one (possibly all) of the co-frames must have $\sigma_i = \sigma_{3DS}$.

A third new parameter in the 3DS grid system is the order of arrangement of the co-frames Z_0, Z_1, Z_2, \dots which can be in any order. We define Z_0 to be the largest scale of elements, Z_1 the next largest, and so on. Thus, a 3-generation 3DS grid system $[X_0, X_1, X_2] = [Z_0, Z_1, Z_2]$ where the co-frames are placed at, $[x_0, x_1, x_2]$ such that $x_0 < x_1 < x_2$, means that the co-frame length scales $[l_0, l_1, l_2]$, are such that $l_0 > l_1 > l_2$. However, the 3DS grid system $[X_0, X_1, X_2] = [Z_1, Z_2, Z_0]$ where the co-frames are placed at, $[x_0, x_1, x_2]$ such that $x_0 < x_1 < x_2$, means that the co-frame length scales $[l_0, l_1, l_2]$, are such that $l_2 > l_0 > l_1$.

It is important to note that σ_{3DS} is much smaller than in the comparative 2DF grid, $\sigma_{3DS} \ll \sigma_{2DF}$. In a 3-generation 3DS system the blockage ratios of the three co-frames is $[\sigma_1, \sigma_2, \sigma_3]$, and with a geometric ratio of $a = 0.5$ between the successive generation, we obtain $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_{2DF}/3$. Therefore, if the blockage ratio of the 2DF is $\sigma_{2DF} = 30\%$, then $\sigma_{3DS} = 10\%$. (This will differ for other $a \neq 0.5$ [8].)

We use Direct Numerical Simulations to compare the mean flows, the turbulence intensities, and the energy spectra generated by three-generation 3DS grid systems for different blockage ratios and different order of co-frame arrangements, and we also compare them to the turbulence produced by RG and 2DF grids. In this study we keep r_1 and r_2 constant. Here, our systems are channels with periodic lateral boundary conditions; the possible effects of no-slip wall conditions and of changing mean flow direction is discussed in Sect. 4, Conclusions.

2 Direct Numerical Simulations

In the first instance we compare the 3DS with the simulations of Laizet et al. [3]. The simulated domain has dimensions of $460.8 \times 115.2 \times 115.2d_{min}^3$ where d_{min} is the thickness of the smallest square. The height and width of the channel is $H = 115.2d_{min}$.

The effective mesh size in the RG is $M_{ef} = 13.33d_{min}$, and the bars have length $115.2d_{min}$, and thickness $2.6d_{min}$. This matches the system reported in [3].

The 2DF has non-dimensionalized lengths and widths $\{l_i, d_i\}$, in generation $i = 0, 1, 2$. Where $l_0 = 57.6 = 0.5h, l_1 = 0.5l_0, l_2 = 0.5l_1$. The bar thicknesses are $d_0 = 8.5, d_1 = 2.9d_2, d_2 = 1$. All lengths are henceforth non-dimensionalized by d_{min} . The time scale is defined by $t_2 = d_{min}/U_\infty$ where U_∞ is the inlet velocity set equal to 1.

The 3DS-2, Table 1, has the same lengths and thickness as the 2DF above, however each generation is held in a co-frame separated from the next by non-dimensional distances, $r_1 = x_1 - x_0 = 17$, and $r_2 = x_2 - x_1 = 8.5$, and $x_0 = 10$, where x_i 's are the non-dimensionalised x -coordinates of the i 'th frame.

The blockage ratio (or solidity) in the RG and 2DF is the same 32%. The maximum blockage ratio in the 3DS is 15%.

OpenFOAM, (Ofoam), was used to create a numerical grid $N_x \times N_y \times N_z = 2304 \times 576 \times 576$. The RG and 2DF grids lie in the plane $x_0 = 10$ downstream of the channel inlet. Periodic boundary conditions were applied on the walls in the y and z directions; and inlet-outlet boundary conditions were applied in the x -direction. The initial condition is a uniform inflow velocity $U_\infty = 1$. The Reynolds number is, $Re = \frac{U_\infty d_{min}}{\nu} = 300$. The resolution is $\Delta x = 0.2d_{min}$ which is adequate for our purposes.

OpenFoam is 2nd order accurate in spatial resolution which is adequate for low Reynolds numbers. It uses finite volume discretization with Pressure Implicit Splitting of Operator Algorithm (PISO). Time discretization using Backward Euler method, whereas gradient and Laplacian term discretization using Gauss linear

Table 1 Different grid types used in this study: the order of arrangement of the co-frames Z_i and the corresponding co-frame blockage ratio σ_i (%) are shown. The last column shows the maximum (i.e. overall) blockage ratio of the grid system

Grid	X_0	σ_0 (%)	X_1	σ_1 (%)	X_2	σ_2 (%)	σ/σ_{3DS} (%)
RG	–	–	–	–	–	–	32
2DF	–	–	–	–	–	–	32
3DS-2	Z_0	15	Z_1	15	Z_2	15	15
3DS-3	Z_0	24	Z_1	15	Z_2	15	24
3DS-4	Z_0	32	Z_1	15	Z_2	15	32
3DS-5	Z_1	15	Z_2	15	Z_0	32	32
3DS-6	Z_2	15	Z_0	32	Z_1	15	32

method are performed. Divergence term discretization is done using Gauss cubic method which is a third order scheme. Interpolation and other terms are discretized using Gauss Linear schemes. The resulting linear systems are solved by preconditioned conjugate gradient method with diagonal incomplete Cholesky preconditioner for pressure solution whereas iterative solver is used with symmetric Gauss-Siedel as the smoother to calculate velocities. Tolerance is set at 10^{-6} . Simulation time step is $\Delta t = 0.015d_{min}/U_\infty$ which corresponds to a Courant number of 0.75. Blockage, such as a bluff body, is achieved by imposing no-slip $u = 0$ condition on the numerical grid corresponding to the surface of the body. The square cross-sectional bars in the 3DS are particularly easy to implement as they match exactly the rectangular geometry of the finite volume elements.

3 Results on Turbulence Intensity

A comparison of the turbulence intensities along different pencils from the RG, 2DF, and 3DS-2 grids from DNS simulations has been reported in [8]. The RG and 2DF plots are close to the results in [3] which validates the DNS for these calculations.

Here, in Fig. 2 we show the time averaged mean flow along the centerline, $U(x = 0)/U_\infty$, from all six grids considered in Table 1, and the centerline time averaged rms turbulence fluctuation (i.e. intensity), $u'(x = 0)/U_\infty$.

Figure 3 shows the turbulence intensity, $u'(x)/U_\infty$, from the same grids along different pencils in the x-direction as indicated.

We group the results into three sets for comparison in Figs. 2 and 3: the first set is (a) and (b), where the 2DF and 3DS-2 are compared. 3DS-2 is obtained from the 2DF by taking the grid bars in 2DF and placing them in the different co-frames.

The second set is (b), (c), and (d), which is a comparison of 3DS grids with different blockage ratios for the same co-frame arrangement, $[Z_0, Z_1, Z_2]$.

The third set is (d), (e), and (f), which is a comparison of 3DS grids with different co-frame arrangements for the same blockage ratio $\sigma_{3DS} = 32\%$.

As expected, for the low blockage ratio 3DS-2 $\sigma_{3DS} = 15\% \ll \sigma_{2DF}$ the mean flow along the centreline in the 3DS-2 is not much disturbed, and the turbulence intensity generated remains low at $\approx 10\%$. However, away from the centreline, the turbulence intensity shows significant peaks in the near field close to the grid, although not as much as in the 2DF. In all cases, in the far field downstream the planar averaged turbulence intensity decays slowly. Thus, it is in the near to mid-range downstream where the differences are most strongly felt.

The results in Figs. 2c and 3c, from the 3DS-3 grid with $\sigma_{3DS} = 24\%$ are remarkably close to the 2DF (32%), Figs. 2a and 3a. The mean and the intensity along the centerline are similar, and the off-centerline turbulence intensities in Fig. 3c display similar trends as well, the peak intensities being only a little higher along most of the pencils.

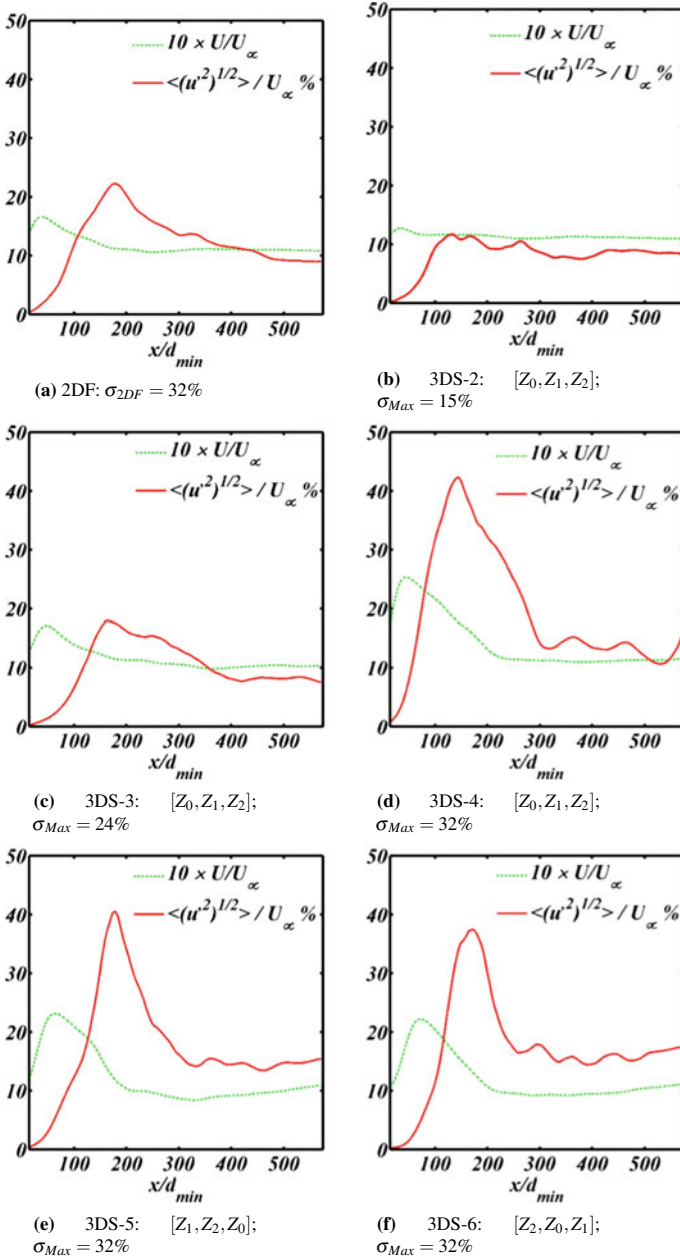


Fig. 2 The mean streamwise velocity U/U_∞ (green), and the streamwise turbulence intensity u'/U_∞ (red) along the centerline, from the 2DF and the 3DS grids. The 3DS co-frame order of arrangement $[Z_i]$, and the blockage ratios σ_{2DF} and σ_{3DF} (%) are shown

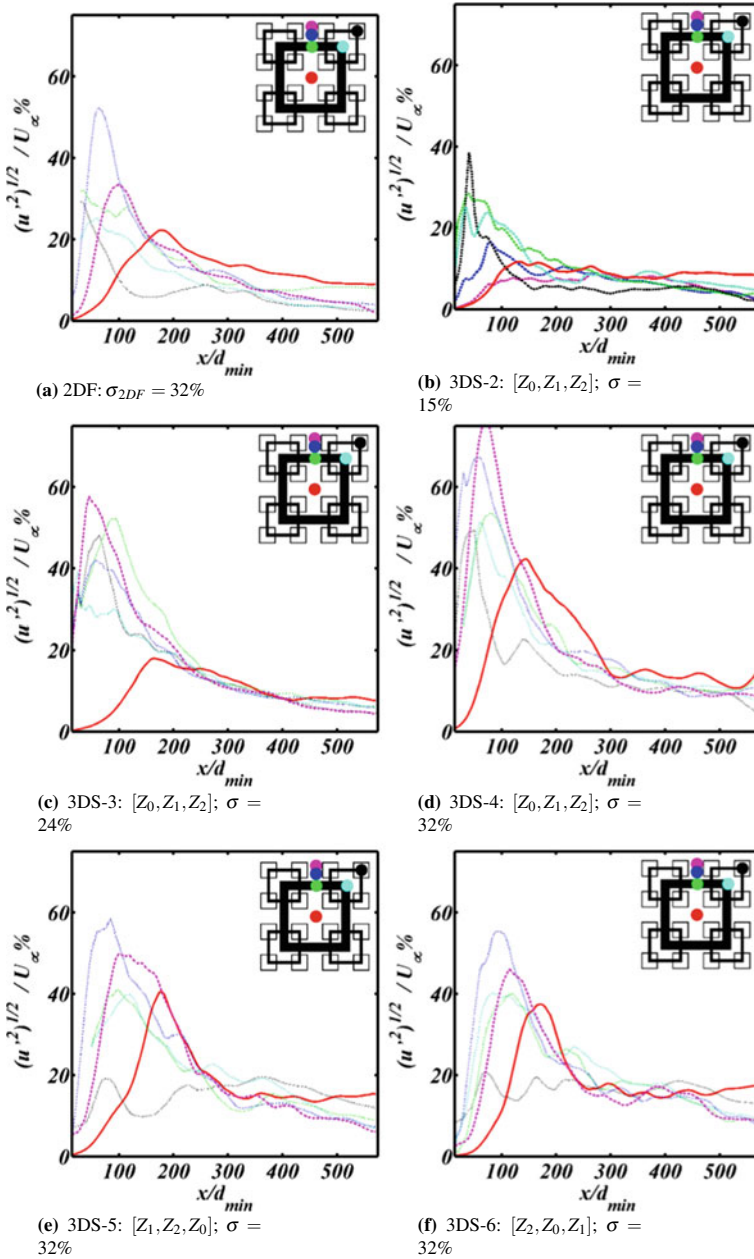


Fig. 3 The streamwise turbulence intensity u'/U_∞ along different pencils as indicated, from the 2DF and the 3DS grids. The 3DS co-frame order of arrangement $[Z_i]$, and the blockage ratios σ_{2DF} and σ_{3DF} (%) are shown

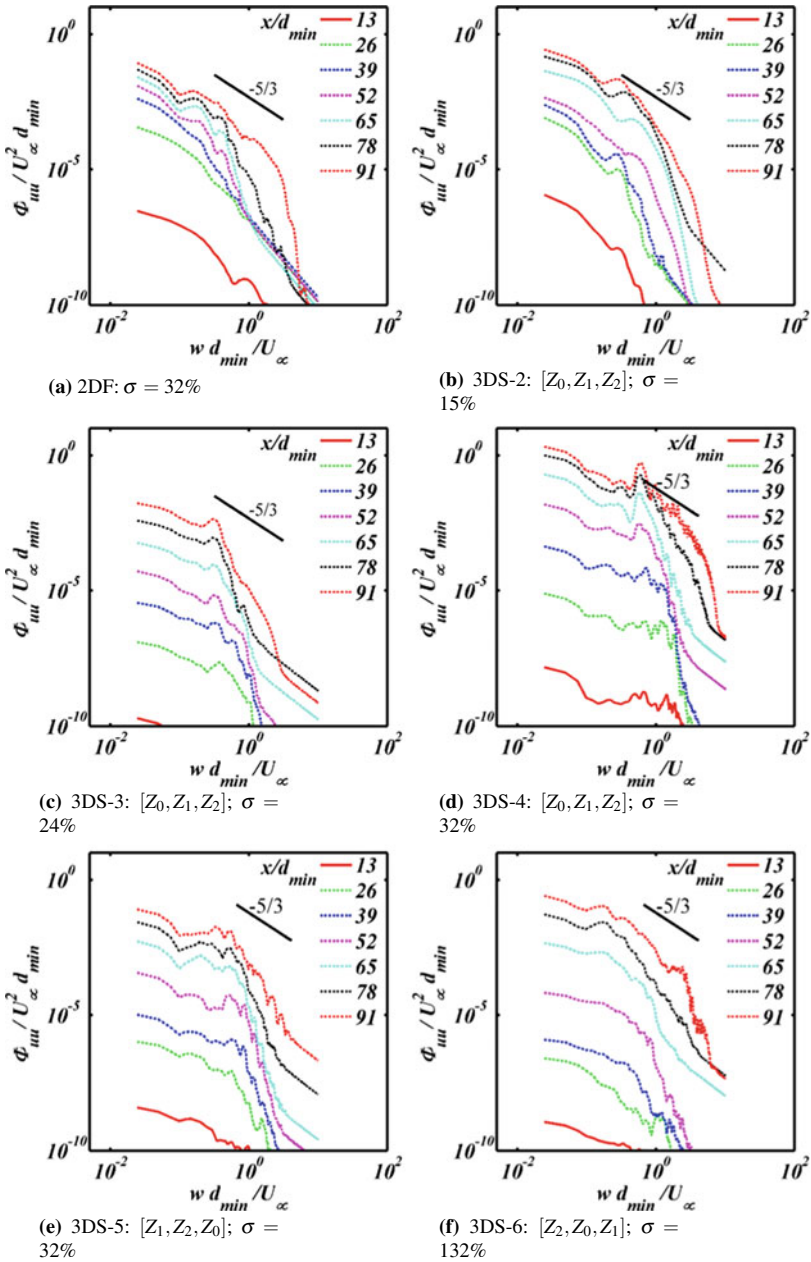


Fig. 4 The energy spectrum $\Phi_{uu}/U_\infty^2 d_{min}$ against the wavenumber $k = wd_{min}/U_\infty$, at different locations along the centerline, from the 2DF and the 3DS grids. The 3DS co-frame order of arrangement $[Z_i]$, and the blockage ratios σ_{2DF} and σ_{3DF} (%) are shown

Figures 2d and 3d, from the 3DS-4 with $\sigma_{3DS} = 32\%$, show the peaks in mean flow and the turbulence intensities exceeding the 2DF peaks by as much as 80% in the near-field downstream. The peaks in Fig. 3(d) are the highest yet observed.

The comparison of the order of arrangement of the co-frames in the 3DS grids for $\sigma_{3DS} = 32\%$, (d)–(f), shows that the turbulence is sensitive to the ordering, although not as sensitive as a change in σ_{3DS} . The three cases are a cyclic permutation, with the largest scale Z_0 being cycled. The order $[Z_0, Z_1, Z_2]$ in the 3DS-4, Fig. 3d, shows the highest peaks in turbulence intensity, although the other two cyclic cases 3DS-5 and 3DS-6 also produce higher peaks than the 2DF. The peaks in the mean flow do not differ much, all three cases being about 50% higher than in the 2DF.

We note that in some of the 3DS grid cases the centerline mean flow and turbulence intensity appear to *increase* far downstream towards the end of the channel. This is almost certainly due to the entrainment of turbulent flows towards the center of the channel, because the current 2DF and 3DS are void of elements in the center. (Other geometric configurations may produce different results.)

Finally, Fig. 4 shows the energy spectrum at different locations downstream for all the grids considered. The spectra are obtained from time series of the velocities at the given location, and converted from frequency domain to the wavenumber domain, $\Phi_{uu}(k)$, where $k \sim wd_{min}/U_\infty$, using Taylor's hypothesis. The 2DF approaches equilibrium turbulence, $\Phi_{uu} \sim k^{-5/3}$ the fastest, and most of the 3DS cases do not achieve this till around $x/d_{min} \approx 100$ remaining in non-equilibrium because the turbulence is still developing in this region. The 3DS appears to prevent a return to equilibrium more effectively than other types of grid.

4 Conclusions

The three-generation 3DS grids that we have investigated show remarkable sensitivity to the blockage ratio σ_{3DS} and the order of arrangement of co-frames when compared to the 2DF grid. Our results show that the three-generation 3DS-3 grid with $\sigma_{3DS} = 24\%$ with co-frame ordering $[Z_0, Z_1, Z_2]$ produces turbulence characteristics that are close to the 2DF with $\sigma_{2DF} = 32\%$; if this could be translated to lower pressure gradient (i.e. lower energy input) then this would be very significant for industrial applications. Furthermore, the sensitivity of the turbulence to the grid parameters implies that a better way of controlling the turbulence generated could be devised. The 3DS grids with blockage ratio equal to the 2DF – 3DS-4, 3DS-5, and 3DS-6, with $\sigma_{3DS} = \sigma_{2DF} = 32\%$ in cyclic co-frame ordering respectively—show peaks in the mean flow and the turbulence intensity in the near field downstream of the grid that greatly exceed that from 2DF grid, by as much as 80% along some pencils. The turbulence spectra show that the turbulence generated by the 3DS grids remain far from equilibrium for the longest period downstream. The entrainment of the turbulence toward the center of the channel causes the mean flow and the intensity to increase far downstream along the centerline.

The results presented here constitute a proof of concept for the 3DS. As this is the first study in 3DS we have simplified the system to facilitate a direct comparison with the RG and 2DF of [3]; we have ignored the boundary wall effects which generates turbulence of its own that would penetrate towards the centre as the streamwise distance increases. However, if the 3DS grid is placed close to the channel entrance, then the effect of boundary walls may not be so important close to the centerline in the near field. It is also of some interest to speculate about how effective the 3DS would be in a bigger system where the mean velocity is changing directions. Shear generated turbulence will likely increase but would need greater pressure drop. On the other hand, if the mixing and turbulence characteristics are dependent mainly on the generation of length scales and time-delay between the co-frames, then it may not matter so much. This and the effect of other parameters, such as varying the inter-frame distances, r_1 and r_2 , is left for future investigation.

Acknowledgements The authors acknowledge the support from King Abdullah University of Science and Technology (KAUST) for making available the High Performance Computing facility Shaheen 2 for this project.

References

1. Queiros-Conde, D., Vassilicos, J.C.: Turbulent wakes in 3-D fractal grids. In: Vassilicos, J.C. (ed.) *Intermittency in Turbulent Flows*. Cambridge University Press (2001)
2. Seoud, R.E., Vassilicos, J.C.: Dissipation and decay of fractal-generated turbulence. *Phys. Fluids* **19**(10), 105108 (2007)
3. Laizet, S., Vassilicos, J.C.: DNS of fractal-generated turbulence. *Flow Turbulence Combust* **87**, 673–705 (2011)
4. Vassilicos, J.C.: Dissipation in turbulent flows. *Ann. Rev. Fluid Mech.* **47**, 95–114 (2015)
5. Sakai, Y., Nagata, K., Suzuki, H., Ito, Y.: Mixing and diffusion in regular/fractal grid turbulence. In: Sakai, Y., Vassilicos, C. (eds.) *Fractal Flow Design: How to Design Bespoke Turbulence and Why*. CISM International Centre for Mechanical Sciences (Courses and Lectures), vol. 568. Springer, Cham (2016)
6. Malik, N.A.: Sparse 3D Multi-Scale Grid Turbulence Generator. US Patent No. US 9,599,269 B2 (2017)
7. Malik, N.A.: Sparse 3D Multi-Scale Grid Turbulence Generator. EPO Patent (2017)
8. Usama, S.M., Tellez-Alvarez, J., Kopec, J., Kwiatkowski, K., Redondo, J.-M., Malik, N.A.: Turbulence Behind 3D Multi-scale Sparse Grids. *IOP Conf. Ser. J. Phys. Conf. Ser.* **1101**, 012048 (2018)

Mathematical Modelling in Engineering, Physical and Chemical Sciences

An Extended Pseudo Potential Multiphase Lattice Boltzmann Model with Variable Viscosity Ratio



Mahmud Ashrafizaadeh, Farshad Gharibi,
and Seyyed Meysam Khatoonabadi

Abstract A new multiphase lattice Boltzmann method (LBM) scheme is proposed through which the viscosity of the two phases can be adjusted based on a theoretical equation of states (EOS). Moreover, any other values of the viscosity can be adjusted by the use of an extra factor of n . The proposed model is validated using two test cases: The Laplace test and the two-phase Poiseuille flow. Numerical results are compared with those of available analytical solutions. A very good agreement between these results are shown. Furthermore, a numerical simulation of a droplet splashing on a thin liquid film is conducted. Despite the standard LBM in which the viscosity of the fluid is bond to the numerical relaxation time and cannot be adjusted, the proposed model enjoys the capability of adjusting the phases' viscosity based on their theoretical and more physically realistic values.

Keywords Lattice Boltzmann method · Pseudo potential · Viscosity ratio · EOS

1 Introduction

Multiphase fluid flow is an important phenomenon in industrial, scientific, and engineering applications. An accurate simulation of multiphase flows is a challenging problem for researchers [3]. An efficient and recently developed computational fluid dynamics method for the simulation of multiphase flows is the lattice Boltzmann method (LBM) [14]. Several LBM multiphase models have been developed within the last three decades [3] (e.g. the Rothman-Keller, the pseudopotential and the free energy models). One of the most prevalent of these models is the pseudopotential model [6] that was proposed by Shan and Chen [15]. However, this model suffers from some deficiencies which limit the applicability of the Shan-Chen model for practical purposes. Some researchers such as Yuan and Schaefer [19] and He et

M. Ashrafizaadeh (✉) · F. Gharibi · S. M. Khatoonabadi
Department of Mechanical Engineering, Isfahan University of Technology (IUT),
8415683111 Isfahan, Iran
e-mail: mahmud@iut.ac.ir

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_69

337

al. [7] incorporated some equations of state (EOS) into the pseudopotential model to achieve higher density ratios. Kupershtokh et al. [10] proposed the exact difference method (EDM) to incorporate the body force into lattice Boltzmann equation to obtain high accuracy in high-density ratios. According to previous studies, most researchers have used the same relaxation times for both phases in the pseudopotential model, i.e. the kinematic viscosity of the phases are equal and the dynamic viscosity ratio is considered equal to the density ratio [4, 8, 18]. According to molecular theory, the dynamic viscosity ratio is proportional to density ratio, but they are not exactly the same [2]. Lou and Li [11] utilized two different relaxation time factors for liquid and gas phases in their pseudopotential multiphase model. In other methods like the free energy model, researchers use a simple linear function of density to change the kinematic viscosity of phases [12, 16]. Chapman and Cowling [2] determined a relation for the viscosity change in terms of the density for dense gases. Suryanarayanan et al. [17] in their lattice Boltzmann multiphase model introduced an expression for viscosity changes by combining the Chapman viscosity relation [2] and the Carnahan-Starling(CS) EOS [10].

This paper aims to link the viscosity of the two phases to the theoretical values calculated by a corresponding EOS as well as to any other desired values which can be adjusted by the use of an extra factor n .

2 Pseudopotential Model

The lattice Boltzmann equation with the Bhatnagar-Groos-Krook (BGK) [1] term and with an external force term can be written as

$$f_i(x + e_i\delta t, t + \delta t) - f_i(x, t) = (1/\tau)[f_i(x, t) - f_i^{eq}(x, t)] + \Delta F_i(x, t) \quad (1)$$

$$\Delta F_i(x, t) = f_i^{eq}(\rho, u + (F_i(x, t)\delta t)/\rho) - f_i^{eq}(\rho, u) \quad (2)$$

where f is the particle distribution function, τ is the relaxation time, which is a function of viscosity by $\nu = (\tau - 0.5)/3$, ΔF is the bulk force, and f_i^{eq} is the equilibrium distribution function as the following [13]

$$f_i^{eq}(x, t) = \omega_i \rho(x, y) \left[1 + \frac{e_i \cdot u}{c_s^2} + \frac{(e_i \cdot u)^2}{c_s^4} + \frac{u^2}{2c_s^2} \right] \quad (3)$$

where ω_i is the i th weighting factor, $C_s = 1/\sqrt{3}$ is the lattice speed of sound, and e_i is the discrete velocity in direction i . Furthermore, ρ and u are the macroscopic density and velocity. When $\delta x = \delta t = 1$, ω_i and e_i for the D2Q9 lattice model are calculated by:

$$e_i = \begin{cases} 0 & \alpha = 0 \\ \cos\left[\frac{(\alpha-1)\pi}{2}\right] \cdot \sin\left[\frac{(\alpha-1)\pi}{2}\right] & \alpha = 1, 2, 3, 4 \\ \sqrt{2}\cos\left[\frac{(\alpha-5)\pi}{2} + \frac{\pi}{4}\right] \cdot \sin\left[\frac{(\alpha-5)\pi}{2} + \frac{\pi}{4}\right] & \alpha = 5, 6, 7, 8 \end{cases} \quad (4)$$

$$\omega_i = [4/9, 1/9, 1/9, 1/9, 1/9, 1/36, 1/36, 1/36, 1/36] \quad (5)$$

After every iteration, the density and momentum are calculated as follows

$$\rho = \sum_i f_i \quad (6)$$

$$\rho u = \sum_i e_i f_i \quad (7)$$

The $F_i(x, t)$ term in Eq. (2) is the inter-particle interaction force among particles of fluid. Kupershtokh et al. suggested a combination of two force terms given by Eq. (8)

$$F_i(x, t) = \left((1 - 2A)\psi(x) \sum_i (\omega_i \psi(x, x') e_i) + A \sum_i (\omega_i \psi^2(x, x') e_i) \right) / \alpha h \quad (8)$$

where h is the lattice spacing, A controls the accuracy of densities compared with the theoretical values, and α is 1.5 in the current settings. $\psi(x)$ is the pseudopotential function. In Eq. (8), $x' = x + e_i$ are the nearest neighbor nodes. In addition, the parameter k is incorporated into the pseudopotential function as

$$\psi = \sqrt{(k\bar{p} - \bar{\rho}c_s^2)} \quad (9)$$

where $\bar{p} = p/p_{cr}$ and $\bar{\rho} = \rho/\rho_{cr}$ that are named non-dimensional pressure and density, respectively, and p_{cr} and ρ_{cr} are the critical pressure and density. The term k is equal to $p_{cr} \Delta t^2 / \rho_{cr} \Delta x^2$. The other parameters, Δt and Δx , are the time step and the lattice spacing, respectively. Generally, $k \approx 0.01$ is an appropriate assumption for a wide range of fluids [10]. The EDM with a non-dimensional CS EOS and a stabilizer parameter k shows a better performance rather than the common form of the EOS [9]

$$\bar{p} = k \left(\frac{c\bar{\rho}R\bar{T}(1 + b\bar{\rho} + (b\bar{\rho})^2 - (b\bar{\rho})^3)}{(1 - b\bar{\rho})^3} - a\bar{\rho}^2 \right) \quad (10)$$

where $\bar{T} = T/T_{cr}$ is the non-dimensional temperature, T is the temperature and T_{cr} is the critical temperature and R is the universal gas constant. Since water is used for present simulations, $a = 3.852462257$, $b = 0.130443884$, and $c = 2.785855166$ are constant as mentioned in reference [10].

In the general pseudopotential model, the viscosity of the two phases cannot be adjusted independently, since the viscosity ratio is equal to the density ratio. Therefore, if one relaxation time is used for both phases (which is a common practice in existing models), the phase kinematic viscosities would be equal. This is apparently not consistent with physical observations. Hence, it is needed to find a theoretical base to physically link the viscosity to the density such that independent kinematic viscosities could be specified for different phases. The coefficient of viscosity in a dense gas (μ) is related to the corresponding coefficient in a normal gas (μ') by Eq. (11) [2].

$$\frac{\mu}{\rho} = b\mu' \left(\frac{1}{b\rho x} + 0.8 + 0.7614b\rho x \right) \quad (11)$$

where $b\rho x$ can be determined from the compressibility of the fluid by $\frac{dP}{dT}$. To couple viscosity and density of the two phases to their theoretical values, it is feasible to utilize an EOS. Since the Carnahan-Starling EOS is utilized in the pseudopotential function, it is used in the following consideration as well

$$\chi^h = b\rho x = \frac{b\rho(1 - \frac{b\rho}{8})}{(1 - \frac{b\rho}{4})^3} \quad (12)$$

By using Eq. (12), one may find the viscosity ratio for different densities of fluid as follows

$$\frac{\mu_1}{\mu_2} = \frac{\rho_1 \left(\frac{1}{\chi_1^h} + 0.8 + 0.7614\chi_1^h \right)}{\rho_2 \left(\frac{1}{\chi_2^h} + 0.8 + 0.7614\chi_2^h \right)} \quad (13)$$

If Eq. (13) is multiplied by a coefficient n , other viscosity ratios could be achieved.

Due to the fact that in the LBM the relaxation time is directly related to the viscosity, the relaxation time for the two phases could be related as follows:

$$\tau_2 = 0.5 + n(\tau_1 - 0.5) \frac{\left(\frac{1}{\chi_1^h} + 0.8 + 0.7614\chi_1^h \right)}{\left(\frac{1}{\chi_2^h} + 0.8 + 0.7614\chi_2^h \right)} \quad (14)$$

Equation (14) allows us to specify different relaxation times for the two phases and as a result their kinematic viscosities would be different, which is the desired outcome.

3 Model Validation

3.1 Static Droplet

In the original pseudopotential model, the relaxation parameter is considered the same for the whole computational domain which results in a fixed viscosity for both phases. However, when Eq. (14) is employed for the relaxation time, each point could have a different viscosity based on local conditions. Figure 1 illustrates relaxation time contours for a static drop surrounded by a different fluid. The present proposed method is used for the simulation. As shown in Fig. 1, there is a distribution of relaxation time (hence viscosity) which changes sharply near the interface due to the sharp density gradient.

In spite of the general claim of the lack of dependency of the relaxation time and the stability of the exact difference method, several numerical results indicate that the EDM simulations become unstable when the relaxation time falls below 0.7. For this reason, the minimum τ should be kept in a particular range. Consequently, it might not be possible to reach the desired values of the viscosity when simulating flows with large density ratios (up to 40). In order to alleviate this problem, the parameter n can be specified to decrease the viscosity ratio as low as possible. Table 1 provides some minimum n values for typical density ratios to ensure a stable simulation. The minimum achievable viscosity ratio of the original pseudopotential model are also compared with those of the present model.

In order to ensure that the proposed model does not have negative side effects, the Laplace’s law was tested. According to the Laplace law, the capillary pressure for a 2D droplet should be such that $P_{in} - P_{out} = \frac{\sigma}{R}$. Where σ is the surface tension and R is the radius of the droplet. In this test, the computational domain ($200 * 200 lu^2$) is filled with a gas phase and a static droplet is placed at the center of the domain.

Fig. 1 Contours of the relaxation time

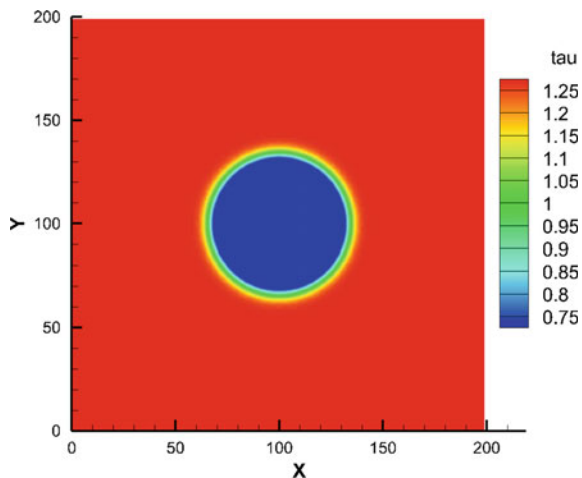
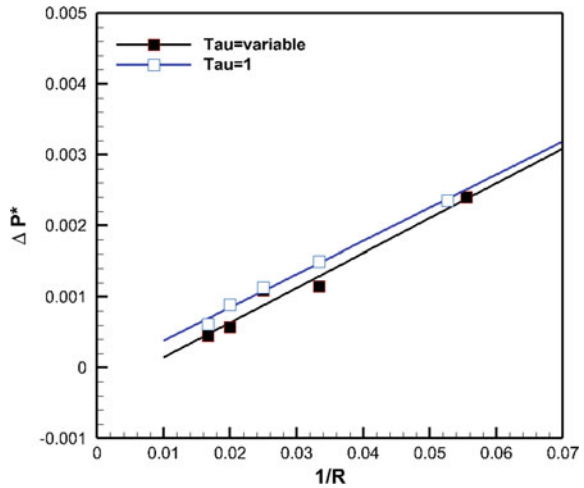


Table 1 Minimum achievable viscosity ratio in different density ratio

Density ratio	26	84	196	530
Viscosity ratio (original model)	26	84	196	530
Min n	1	5	8	18
Min achievable viscosity ratio	26.21	30.44	82.64	221.76

Fig. 2 Pressure difference as a function of inverse drop radius

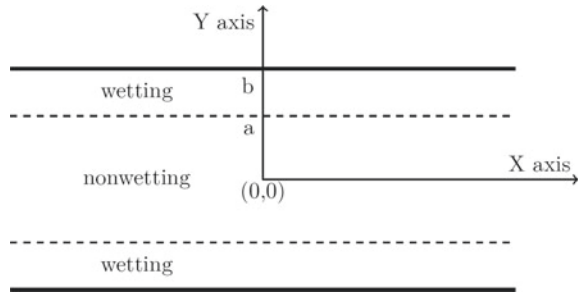


By changing the radius of the droplet, the pressure difference between the inside and outside of the droplet is obtained. Figure 2 shows the pressure difference versus the inverse of the radius so that the slope of the plot shows the computed surface tension. The results demonstrate that the surface tension and the viscosity ratio in the present model are independent, as it should be expected. It is worthwhile mentioning that the magnitude of the spurious current does not change noticeably when Eq. (14) is employed.

3.2 Immiscible Two-Phase Poiseuille Flow

One of the two-phase flows that has an analytical solution is the co-current multi-phase flow between two parallel flat walls. As shown in Fig. 3, the wetting phase has contact with the walls, and the non-wetting phase flows between the films of the wetting phase. In this problem, the velocity of the non-wetting phase is affected by the dynamic viscosity ratio of the fluids [18], $M = \frac{\mu_{nw}}{\mu_w}$. In most previous LBM simulations, the kinematic viscosity of both fluids are considered to be equal

Fig. 3 Schematic of Immiscible two-phase poiseuille flow



(by using a fixed relaxation time). Hence, the dynamic viscosity ratio would be equal to the density ratio i.e. $M = \frac{\mu_{nw}}{\mu_w} = \frac{\rho_{nw}}{\rho_w}$ [4, 8, 18]. But in reality, for many fluids this is not a physical accurate assumption. In the present model, however, the dynamic viscosity ratio can be adjusted as a function of the density ratio, allowing us to have different relaxation times (i.e. kinematic viscosities) for different phases.

By assuming a 2D Poiseuille flow between two flat plates, the analytical velocity profile is [18]:

$$u = \frac{F}{2\mu_w}(L^2 - y^2) \quad a < |y| < L \tag{15}$$

$$u = \frac{F}{2\mu_w}(L^2 - a^2) + \frac{F}{2\mu_{nw}}(a^2 - y^2) \quad 0 < |x| < a \tag{16}$$

where μ_w and μ_{nw} are dynamic viscosities of the wetting and non-wetting phases. F is the body force that is equal to the pressure gradient i.e. $\frac{\Delta P}{l}$. The relative permeability of each phase is defined by [18]:

$$K_{r.w} = \frac{1}{2}S_w^2(3 - S_w) \quad K_{r.nw} = S_{nw}[\frac{3}{2}M + S_{nw}^2(1 - \frac{3}{2}M)] \tag{17}$$

where $K_{r.w}$ and $K_{r.nw}$ are the wetting and non-wetting relative permeabilities, respectively. S_w and S_{nw} are the wetting and non-wetting saturations. The saturation of each phase can be defined as the occupied width of the one phase divided by the entire width of the channel [5].

In this study, a $120 * 240 l u^2$ computational domain with periodic boundary conditions in the x-direction and bounce-back boundary conditions for the upper and lower walls are considered. A constant external force is imposed on both phases to move the flow inside the channel. Figure 4 shows the velocity profiles of the two-phase flow at different wetting saturations. In this case, a gas density of 0.102, and a fluid density of 2.57 is assumed. As illustrated in Fig. 4, the present LBM results show good agreement with those of the analytical solution. However, by increasing the saturation of the non-wetting phase, a small deviation from analytical results are detected. This difference originates from the effect of the interface thickness on the calculated velocities. In fact, when the wetting layer's height is in the order of

Fig. 4 Velocity profile perpendicular to the direction of the flow for different saturation

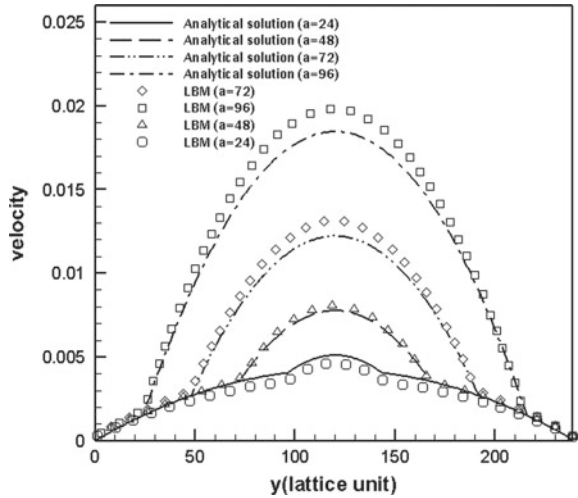
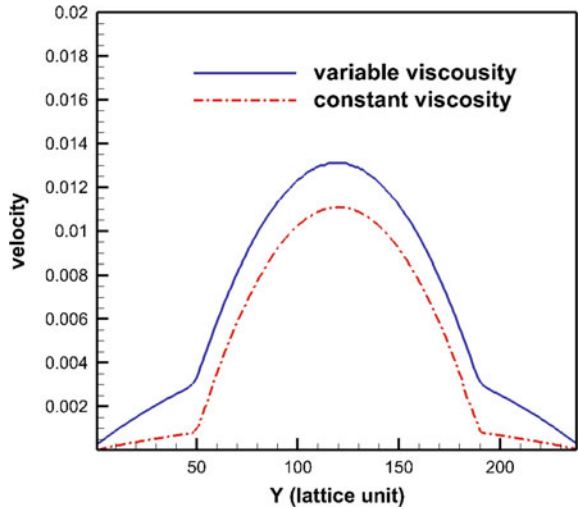


Fig. 5 Velocity profile perpendicular to the direction of the flow for original and new method in a same condition



the interface thickness, the velocity error at the interface affects the total calculated velocity of the wetting phase. Figure 5 compares the velocity profiles across the channel for $a = 72$ for the present model and the original model. As shown, the original model under predicts the velocity profile.

Using the calculated velocity profiles, the relative permeability of both phases can be determined as a function of the saturation and the viscosity ratio. Figure 6 shows the relative permeability of the wetting phase calculated by both the general pseudopotential model and the new model. According to Eq. (17), the relative permeability of the wetting phase does not depend on the viscosity changes of the phases. Based on Fig. 6, it can be seen that the proposed model illustrates this point well.

Fig. 6 Relative permeabilities for the wetting phases

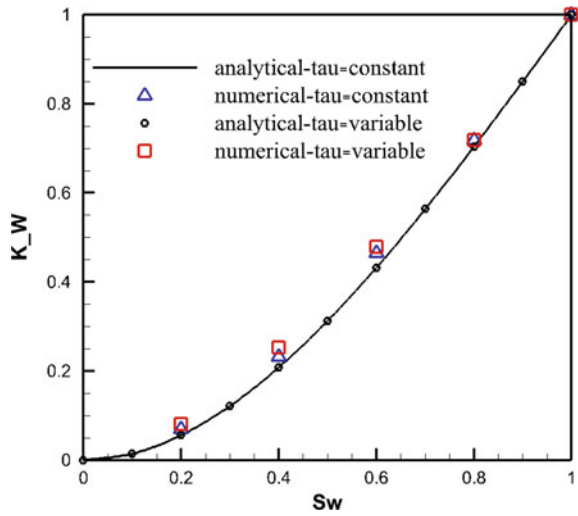


Fig. 7 Relative permeabilities for the non-wetting phases

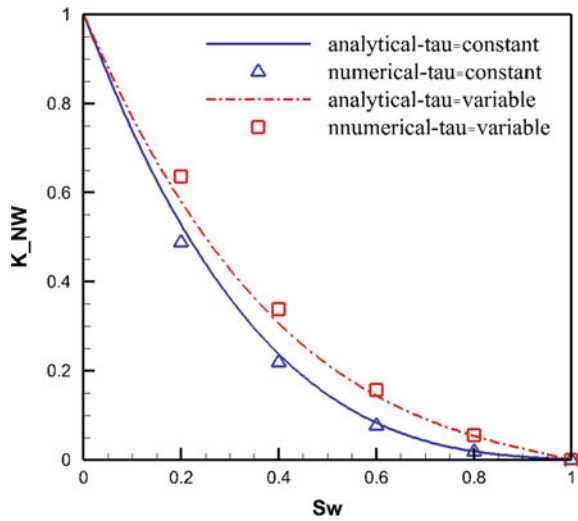


Figure 7 also illustrates the relative permeability of the non-wetting phase, where the effect of changing the viscosity of the phases can be observed. The results show that there is a good agreement between the LBM simulation and that of the analytical solution.

3.3 Droplet Splashing

In this section, the problem of a droplet, with an initial velocity, splashing on a thin liquid film is investigated. The effect of using a variable viscosity model is compared with that of a constant viscosity assumption. Figure 8 shows an overview of the initial conditions of the problem. For the simulation, a drop with a radius of 50 (lattice unit) and an impact velocity of $V = -0.13$ is considered in a $250 * 800$ lattice unit solution domain. The reduced temperature is set to be $\bar{T} = 0.5$, and the relaxation time of the gas phase is set to be $\tau = 1.8$.

Figures 9 and 10 show the snapshots of this simulation at different times. As can be seen from Fig. 9, in constant viscosity model, only an outward wave is generated whereas in the modified present model, as shown in Fig. 10, a crown is created radially away from the center, which is qualitatively more consistent with actual observations.

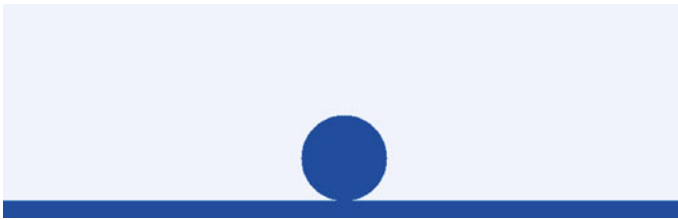


Fig. 8 Initial condition of droplet splashing on a thin liquid film (dark blue = liquid phase, light blue = gas phase)

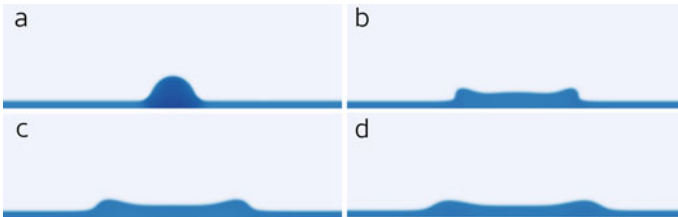


Fig. 9 Density contours of the splashing simulation by constant viscosity model at **a** $t = 300$, **b** $t = 1200$, **c** $t = 2100$, **d** $t = 3000$ (lattice unit)

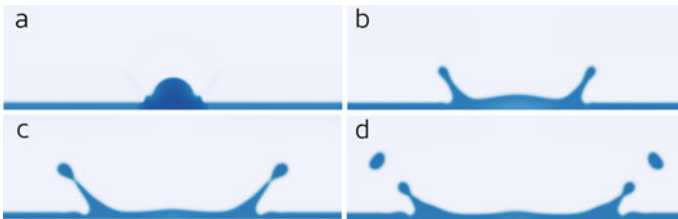


Fig. 10 Density contours of the splashing simulation by variable viscosity model at **a** $t = 300$, **b** $t = 1200$, **c** $t = 2100$, **d** $t = 3000$ (lattice unit)

4 Conclusion

In this study, a new scheme is proposed for modeling multiphase flows through which viscosity of two phases can be adjusted based on a theoretical equation. In order to show the capabilities of the suggested model, the Laplace test has been conducted. The results illustrate that the density and the surface tension are independent of the specified viscosity. Also, the relative permeability tests for a two-phase flow between parallel plates, show a good agreement with those of analytical results. With the decrease of the wetting fluid saturation, the simulated velocity profile slightly deviated from that of the analytical values. Moreover, a coefficient n is incorporated into the viscosity equation that can be utilized to specify arbitrary viscosity ratios. However, this factor is limited, especially when large density ratios are considered. The present model seems to be a promising tool to play a significant role in the simulation of practical applications, such as phase transition flows.

References

1. Bhatnagar, P.L., Gross, E.P., Krook, M.: A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. *Phys. Rev.* **94**(1–12), 511–525 (1954)
2. Chapman, S., Cowling, T.G.: *The Mathematical Theory of Non-uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases*. Cambridge University Press (1970)
3. Chen, L., Kang, Q., Mu, Y., He, Y.-L., Tao, W.-Q.: A critical review of the pseudopotential multiphase lattice Boltzmann model: methods and applications. *Int. J. Heat Mass Transf.* **76**, 210–236 (2014)
4. Dou, Z., Zhou, Z.-F.: Numerical study of non-uniqueness of the factors influencing relative permeability in heterogeneous porous media by lattice Boltzmann method. *Int. J. Heat Fluid Flow* **42**, 23–32 (2013)
5. Ghassemi, A., Pak, A.: Numerical study of factors influencing relative permeabilities of two immiscible fluids flowing through porous media using lattice Boltzmann method. *J. Pet. Sci. Eng.* **77**(1), 135–145 (2011)
6. Haibo Huang, M.S., Lu, X.-Y.: *Multiphase Lattice Boltzmann Methods: Theory and Application*. Wiley, June 2015
7. Hu, A., Li, L., Chen, S., Liao, Q., Zeng, J.: On equations of state in pseudo-potential multiphase lattice Boltzmann model with large density ratio. *Int. J. Heat Mass Transf.* **67**, 159–163 (2013)
8. Huang, H., Li, Z., Liu, S., Lu, X.-Y.: Shan-and-Chen-type multiphase lattice Boltzmann study of viscous coupling effects for two-phase flow in porous media. *Int. J. Numer. Methods Fluids* **61**(3), 341–354 (2009)
9. Khatoonabadi, S.M., Ashrafizaadeh, M.: Comparison and development of multiphase pseudo-potential model for various equations of state. *Modares Mech. Eng.* **15**(12), 376–386 (2015) (in Persian)
10. Kupershtokh, A., Medvedev, D., Karpov, D.: On equations of state in a lattice Boltzmann method. *Comput. Math. Appl.* **58**(5), 965–974 (2009)
11. Li, Q., Luo, K.H., Li, X.J.: Lattice Boltzmann modeling of multiphase flows at large density ratio with an improved pseudopotential model. *Phys. Rev. E* **87**, 053301 (2013)
12. Liu, H., Zhang, Y.: Droplet formation in a T-shaped microfluidic junction. *J. Appl. Phys.* **106**(3), 034906 (2009)

13. McNamara, G.R., Zanetti, G.: Use of the Boltzmann equation to simulate lattice-gas automata. *Phys. Rev. Lett.* **61**(20), 2332 (1988)
14. Sbragaglia, M., Benzi, R., Biferale, L., Succi, S., Sugiyama, K., Toschi, F.: Generalized lattice Boltzmann method with multirange pseudopotential. *Phys. Rev. E* **75**, 026702 (2007)
15. Shan, X., Chen, H.: Lattice Boltzmann model for simulating flows with multiple phases and components. *Phys. Rev. E* **47**(3), 1815 (1993)
16. Shi, Y., Tang, G.H.: Lattice Boltzmann simulation of droplet formation in non-Newtonian fluids. *Commun. Comput. Phys.* **17**(4), 1056–1072 (2015)
17. Suryanarayanan, S., Singh, S., Ansumali, S.: Extended BGK Boltzmann for dense gases. *Commun. Comput. Phys.* **13**(3), 629–648 (2013)
18. Yiotis, A.G., Psihogios, J., Kainourgiakis, M.E., Papaioannou, A., Stubos, A.K.: A lattice Boltzmann study of viscous coupling effects in immiscible two-phase flow in porous media. *Colloids Surf. A: Physicochem. Eng. Asp.* **300**(1), 35–49 (2007)
19. Yuan, P., Schaefer, L.: Equations of state in a lattice Boltzmann model. *Phys. Fluids* **18**(4), 042101 (2006)

Approximating Dispersive Materials with Parameter Distributions in the Lorentz Model



Jacqueline Alvarez, Andrew Fisher, and Nathan L. Gibson

Abstract We seek to improve the accuracy of the Lorentz model by incorporating a distribution of dielectric parameters and introducing a microscopic quantity we call the random polarization. Thus the usual polarization is the macroscopic average, or expected value, of the random polarization. The forward problem in the frequency domain demonstrates the difference between the distributed and deterministic models. Using a least squares cost formulation and χ^2 significance test, we explore the parameter identification problem for saltwater data. For analysis in the time domain, we use Polynomial Chaos and the Finite Difference Time Domain methods to discretize in one dimension. We then examine two time domain inverse problems that compare interrogation signals.

Keywords Lorentz model · Random polarization · Polynomial chaos · Parameter estimation

1 Introduction

Electromagnetic interrogation of dispersive materials is of current interest in industry for its potential as a non-invasive method in identifying weaknesses or compositions in materials. An example is determining a material's dispersive properties through the analysis of a single transmitted ultra-wideband (UWB) pulse. Several different methods have been suggested that expand on the common Lorentz polarization

J. Alvarez
University of California, Merced, Merced, USA
e-mail: jalvarez94@ucmerced.edu

A. Fisher
University of California, Los Angeles, USA
e-mail: andrew.fisher@physics.ucla.edu

N. L. Gibson (✉)
Oregon State University, Corvallis, USA
e-mail: gibsonn@math.oregonstate.edu

model, some employing linear combinations of poles or normally distributed poles to fit models to data [5]. In this paper, however, we explore placing beta distributions on the dielectric parameters in the model.

First we present necessary background information including Maxwell's equations, the constitutive equations, and the Lorentz model. Next we introduce random parameters and define random polarization. Then using Fourier transforms, we explore the frequency domain through the complex permittivity and present a parameter identification problem. For analysis in the time domain, we use Polynomial Chaos and the Finite Difference Time Domain (FDTD) method to discretize in one dimension. We then examine two time domain inverse problems that compare interrogation signals.

2 Background

2.1 Maxwell's Equations

We begin by presenting Maxwell's equations that describe the behavior of electromagnetic waves in free space. D is the electric flux density, E and H are the electric and magnetic fields. The magnetic permeability of free space is given by μ_0

$$\frac{\partial D}{\partial t} + J = \nabla \times H \quad (1a)$$

$$\mu_0 \frac{\partial H}{\partial t} = -\nabla \times E \quad (1b)$$

$$\nabla \cdot D = 0 \quad (1c)$$

$$\nabla \cdot B = 0 \quad (1d)$$

Next, we incorporate the constitutive laws that adapt Maxwell's equations for propagation in materials. We let $\tilde{\epsilon}$ represent the electric permittivity which is equal to the product of the permittivity of free space and a relative permittivity ($\tilde{\epsilon} = \epsilon_0 \epsilon_\infty$). The polarization in the material is given by P , defined by

$$D = \tilde{\epsilon} E + P. \quad (2)$$

To find the equations defining electromagnetic waves in a material, we substitute the constitutive equations into Maxwell's curl equations and reduce to one dimension:

$$\tilde{\epsilon} \frac{\partial E_x}{\partial t} = -\frac{\partial H_y}{\partial z} - \frac{\partial P_x}{\partial t} \quad (3)$$

$$\mu_0 \frac{\partial H_y}{\partial t} = -\frac{\partial E_x}{\partial z}. \quad (4)$$

From now on, we drop the subscripts so that $E(t, z) = E_x(t, z)$, $P(t, z) = P_x(t, z)$, and $H(t, z) = H_y(t, z)$.

Prior to interrogation, there are no fields or polarizations present so our initial conditions are:

$$E(0, z) = H(0, z) = P(0, z) = 0. \quad (5)$$

Our boundary conditions include the interrogating signal, $f(t)$, at $z = 0$ and a reflective surface at $z = z_0$:

$$E(t, 0) = f(t) \text{ and } E(t, z_0) = 0. \quad (6)$$

2.2 Lorentz Model

There are several models that describe polarization in materials. In this paper, we focus on the Lorentz model [4] for which the physical assumption is that we can treat electrons in the material as simple harmonic oscillators. The Lorentz model is given by

$$\ddot{P} + 2\nu\dot{P} + \omega_0^2 P = \varepsilon_0 \omega_p^2 E \quad (7)$$

where ν is the damping coefficient, ω_0 is the natural resonant frequency and ω_p is the plasma frequency.

Taking the Fourier transform of (2) [4], we get $\hat{D} = \varepsilon_0 \varepsilon(\omega) \hat{E}$ where $\varepsilon(\omega)$ is called the *complex permittivity*, and is given by

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\omega_p^2}{\omega_0^2 - \omega^2 + i2\nu\omega}. \quad (8)$$

It is useful to separate (8) into its real and imaginary parts. The real part is primarily responsible for the material effect on the frequency dependent speed of wave propagation, or dispersion, while the imaginary part is primarily responsible for loss or dissipation. As such, it is common to write the imaginary part as an effective (frequency dependent) conductivity. Thus the separation is $\varepsilon(\omega) = \varepsilon_r(\omega) - i\sigma(\omega)/\omega$:

$$\varepsilon_r = \varepsilon_\infty + \frac{\omega_p^2(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + 4\nu^2\omega^2} \quad (9a)$$

$$\sigma = \frac{\omega_p^2 2\nu}{(\omega_0^2 - \omega^2)^2 + 4\nu^2\omega^2}. \quad (9b)$$

2.3 Random Polarization

In this paper, we explore the effects of altering the original Lorentz model by applying a probability distribution to the resonance frequency, or rather $\eta = \omega_0^2$ (since ω_0 always appears as ω_0^2 , we choose to vary ω_0^2 for simplicity). In order to use distributions of parameters with Maxwell's equations [8], we define the random Lorentz model similar to (7), but where the resonance frequency is now a random variable and \mathcal{P} is the *random polarization*:

$$\ddot{\mathcal{P}} + 2\nu\dot{\mathcal{P}} + \eta\mathcal{P} = \varepsilon_0\omega_p^2 E. \quad (10)$$

Next, we declare that the macroscopic polarization P , defined in (2), is modeled by the expected value of the random polarization, where η is a random variable defined over $[a,b]$ with probability density function $f(\eta)$ [2]:

$$P(t, z) = \mathbb{E}[\mathcal{P}] := \int_a^b \mathcal{P}(t, z; \eta) f(\eta) d(\eta). \quad (11)$$

For example, in the case of a uniform probability distribution, $f(\eta) = \frac{1}{b-a}$.

Thus, (11) along with (10) represents a more sophisticated model for the macroscopic polarization present in a material than does the simple Lorentz model given in (7). We note that the Lorentz model is a subset of the random Lorentz model which assumes a discrete distribution of parameters consisting of a single value.

3 Frequency Domain

Now we consider the frequency domain formulation of the random Lorentz model. The complex permittivity (8) becomes

$$\varepsilon(\omega; \eta) = \varepsilon_\infty + \frac{\omega_p^2}{\eta - \omega^2 + i2\nu\omega}. \quad (12)$$

An observed, measured value for the permittivity or conductivity of a material would represent a macroscopic average of a microscale phenomenon. Thus, in order to compare this random Lorentz model for complex permittivity to data, we must compute its expected value. Because η is a random variable, we must integrate over the corresponding probability distribution to find the expected complex permittivity. In the case of a uniform distribution, it turns out that there is an analytical formula [1]. Otherwise, numerical quadrature can be used.

3.1 Frequency Domain Inverse Problem

The complex permittivity describes how a signal will propagate in a Lorentz material with given parameters. The frequency domain inverse problem involves recovering the appropriate material parameters, say q , by fitting a model to experimental data. Using a least squares cost formulation, we optimize using the MATLAB `lsqnonlin` function. We consider a fixed range of frequencies, and a uniform mesh on this range. We assume that permittivity and conductivity measurements are available corresponding to these discrete frequencies. We let the permittivity and conductivity data vectors be concatenated into a single vector V_{data} . Then, given a trial set of material parameters, q , a complex permittivity model (either the deterministic Lorentz or the random Lorentz) can be evaluated at the same discrete frequencies and will produce a vector of complex permittivity values $V_{model}(q)$ to compare to the data. The residual ($R(q)$) of this process is defined as the difference between the measured data and the model estimate. The least squares cost ($F(q)$) is defined as the norm of the residual, thus

$$R = V_{data} - V_{model} \tag{13}$$

$$F = R^T R. \tag{14}$$

If the permittivity and conductivity are on the same order of magnitude, they do not need to be scaled relative to each other. Thus our parameter estimation problem is to find q such that $F(q)$ is minimized.

We want to show that random permittivities are distinct from deterministic permittivities. For example, [9] discusses how the Lorentz-Lorentz model for permittivity is actually equivalent to the shifted Lorentz model with equivalence when the inequality $\omega_p^2 \ll 6\nu\omega_0$ is satisfied. To be sure the permittivities are distinct, we apply a deterministic fit of parameters to data which comes from a uniform distribution. For comparison, we plot the deterministic permittivity, i.e., using the expected value with no distribution. Results are shown in Fig. 1 where the distribution’s (relative) range is the radius divided by the midpoint ($r = \frac{b-a}{b+a}$). As expected, the deterministic permittivity model was unable to fit the random permittivity data.

We now attempt to fit parameters to actual saltwater data from [10]. The fits and results are shown in Fig. 2 and Table 1. The error in the deterministic model fit is twice that of the distributional model fit.

To determine if there is statistical significance between the fits, we use the hypothesis testing presented in [3]. First we let $q = (\nu, \mathbb{E}[\omega_0^2], \omega_p, r) \in Q$ where Q is the parameter set. Then, we define Q_0 to be the set $\{Q_0 \in Q : r = 0\}$ (e.g., a discrete distribution, or a deterministic model) and let \hat{q}_ℓ and \bar{q}_ℓ denote minimizers of Q_0 and Q , respectively. We construct the hypotheses $H_0 : r = 0$ and $H_A : r \neq 0$ so that a rejection of the null hypothesis correlates to a difference in the fits. Finally, we define the test statistic:

$$U_\ell = \frac{\ell [F_\ell(\hat{q}_\ell) - F_\ell(\bar{q}_\ell)]}{F_\ell(\bar{q}_\ell)} \tag{15}$$

Table 1 Results for saltwater fits

Source	ε_∞	$\nu (1 \times 10^{13})$	$\omega_0 (1 \times 10^{14})$	Range	$\omega_p (1 \times 10^{14})$	Cost
Det. fit	1.7931	2.7547	6.3568	–	1.7333	0.1704
Dist. fit	1.7901	1.6112	6.3608	0.0855	1.6067	0.0655

where ℓ is the number of data points and F_ℓ is the minimized cost.

We proceed by using a significance level α and $\chi^2(s)$ distribution with s degrees of freedom to obtain the threshold τ so that $P(\chi^2(s) > \tau) = \alpha$. We compare U_ℓ with τ , such that if $U_\ell > \tau$ we reject the null hypothesis H_0 . Because the parameter r is the only degree of freedom ($s = 1$), we refer to Table 2.

Our simulations return $F_\ell(\hat{q}) = 0.1704$ and $F_\ell(\bar{q}) = 0.0655$ with $\ell = 79$. Plugging those values into (15) we get $U_\ell = 126.584$. Because $U_\ell \gg \tau$, we reject H_0 . Thus, we can conclude that a distributed model provides a statistically significantly better fit than a deterministic model.

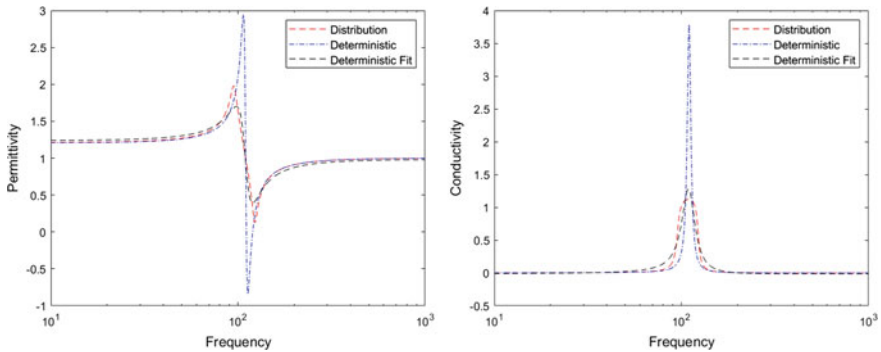


Fig. 1 Plots of the permittivity data, as well as the fitted model, for synthetic data using parameters: $\nu = 3$, $\omega_p = 50$, $\mathbb{E}[\omega_0^2] = 110$, and $r = 0.25$

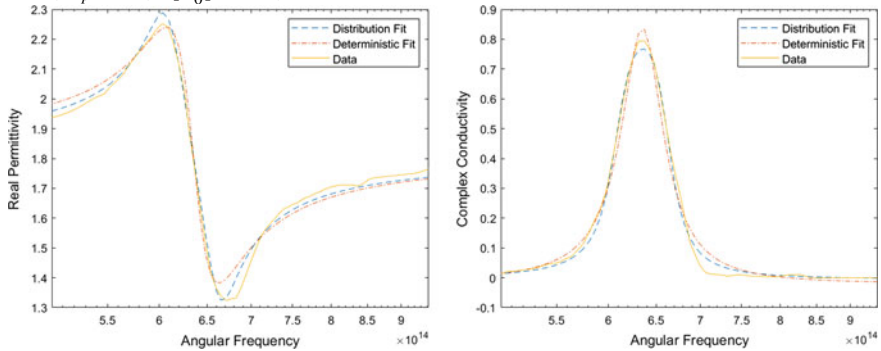


Fig. 2 Plots of the permittivity data, as well as the fitted model, for saltwater data

Table 2 χ^2 distribution with 1 degree of freedom

$\alpha = 0.25$	$\tau = 1.32$
$\alpha = 0.10$	$\tau = 2.71$
$\alpha = 0.05$	$\tau = 3.84$
$\alpha = 0.01$	$\tau = 6.63$
$\alpha = 0.001$	$\tau = 10.83$

Table 3 Bimodal fit comparison

Source	ε_∞	ν	ω_0	Range	ω_p	ν_2	$\omega_{0,2}$	Range	$\omega_{p,2}$	Cost
Data	1.000	13.000	110.000	0.200	50.000	20.000	150.000	0.300	70.000	–
Uni-modal	0.986	14.659	134.811	0.539	83.861	–	–	–	–	4.763
Bi-modal	0.978	15.079	110.918	0.179	53.817	20.573	151.327	0.262	68.229	0.1118
Bi-discrete	0.970	17.693	111.580	–	55.571	27.731	151.144	–	71.073	0.4894

3.2 Bimodal Data

We also consider fitting parameters to bimodal data. First, we create data using a distribution with the parameters given in Table 3. Because real data requires repeated measurements, instrument errors can be propagated. For this reason, we add normally distributed noise with $\mu = 0$ and $\sigma = 0.001$ to the derivatives of the bimodal data. Then we optimize with uni-modal, bi-modal, and bi-discrete model fits. Results are given in Table 3. As expected, the bi-modal model fit best matches the data with $F = 0.1118$, the uni-modal cost was 10 times larger.

4 Time Domain Discretization

Now we consider the time domain formulation of the random Lorentz model, using Polynomial Chaos to deal with the random variable ω_0^2 . Polynomial Chaos is a method of solving random differential equations by expressing quantities as orthogonal polynomial expansions in the random variable [11]. We expand in the normalized Jacobi polynomials, but because they are defined on $[-1, 1]$ it is necessary to scale our distribution. Letting $\omega_0^2 = m + r\xi$ so that ξ is defined on $[-1, 1]$, we identify m and r as the center and radius of the distribution. Random polarization can now be expressed as a function of ξ ,

$$\mathcal{P}(\xi, t) = \sum_{i=0}^{\infty} \alpha_i(t) \phi_i(\xi). \tag{16}$$

We refer the reader to the details in [1, 7], which include a rigorous analysis of the stability and dispersion properties of the FDTD method described here. Each of these is an extension of the methods developed in [8].

Letting $\dot{\alpha} = \beta$ we express the polynomial chaos modal equations for (10) as a system of differential equations:

$$\dot{\alpha} = \beta \tag{17a}$$

$$\dot{\beta} = -A\alpha - 2\nu I\beta + \mathbf{f}, \tag{17b}$$

where $\mathbf{f} = \hat{\epsilon}_1 \epsilon_0 \omega_p^2 E$.

4.1 FDTD Discretization

Combining Maxwell’s equations with our results from Polynomial Chaos, we have the four equations that completely determine propagation through a dielectric material. We repeat them here as a reference:

$$\epsilon_\infty \epsilon_0 \frac{\partial E}{\partial t} = -\frac{\partial H}{\partial z} - \beta_0 \tag{18a}$$

$$\frac{\partial H}{\partial t} = -\frac{1}{\mu_0} \frac{\partial E}{\partial z} \tag{18b}$$

$$\dot{\alpha} = \beta \tag{18c}$$

$$\dot{\beta} = -A\alpha - 2\nu I\beta + \mathbf{f}. \tag{18d}$$

It is important to note that $\frac{\partial P}{\partial t}$ is the time change in macroscopic polarization or the time change of the expected value of our random polarization. Since only the 0th Jacobi polynomial is constant, we identify $\beta_0 = \frac{\partial P}{\partial t}$ with the other polynomials and coefficients determining uncertainties. This explains our substitution in (18a).

To model these equations, we discretize them according to the one-dimensional Yee Scheme [12]. The Yee Scheme implements a staggered grid where the electric field and random polarization are evaluated at integer time steps and spatial steps, while the magnetic field is evaluated at half integer time steps and spatial steps. We consider the domain $z \in [0, z_0]$ for $t \in [0, T]$, choosing integers J and N to discretize so that $\Delta z = \frac{z_0}{J}$ and $\Delta t = \frac{T}{N}$. Let $z_j = j\Delta z$ and $t^n = n\Delta t$. If U is a field variable, we define the grid function to be

$$U_j^n \approx U(x_j, t^n).$$

Our discrete initial conditions and boundary conditions are:

$$E_j^0 = \alpha_j^0 = \beta_j^0 = 0 \text{ for } 0 \leq j \leq J, \quad H_j^n = 0 \text{ for } 0 \leq j \leq J \text{ and } n \leq 0,$$

$$E_0^n = f(t^n) \text{ and } E_j^n = 0 \text{ for } 0 \leq n \leq N.$$

First we approximate derivatives with finite differences and constant terms with averages:

$$\varepsilon_\infty \varepsilon_0 \frac{E_j^{n+1} - E_j^n}{\Delta t} = -\frac{H_{j+\frac{1}{2}}^{n+\frac{1}{2}} - H_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta z} - \frac{\beta_{0,j}^{n+1} + \beta_{0,j}^n}{2} \quad (19a)$$

$$\frac{H_{j+\frac{1}{2}}^{n+\frac{1}{2}} - H_{j+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} = -\frac{1}{\mu_0} \frac{E_{j+1}^n - E_j^n}{\Delta z} \quad (19b)$$

$$\frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} = \frac{\beta_j^{n+1} + \beta_j^n}{2} \quad (19c)$$

$$\frac{\beta_j^{n+1} - \beta_j^n}{\Delta t} = -A \frac{\alpha_j^{n+1} + \alpha_j^n}{2} - 2\nu I \frac{\beta_j^{n+1} + \beta_j^n}{2} + \frac{\hat{\varepsilon}_1 \varepsilon_0 \omega_p^2}{2} [E_j^{n+1} + E_j^n]. \quad (19d)$$

Equations (19a), (19c), and (19d) are defined for $\{1 \leq j \leq J - 1, 0 \leq n \leq N - 1\}$ and (19b) is defined for $\{0 \leq j \leq J - 1, 0 \leq n \leq N - 1\}$.

5 Time Domain Inverse Problem

In this section, we apply our forward simulation to the time domain inverse problem. It was proven in [2] that unique solutions exist for time-domain parameter identification problems involving dispersive Maxwell's equations posed with distributions over dielectric parameters. Specifically, we wish to reconstruct the parameters of a material from noisy data collected by a receiver a distance of $0.252 \mu\text{m}$ into the material. We borrow parameter values from [4]. Assuming $\varepsilon_\infty = 1$ and that the interrogating signal is known, only three parameters need to be optimized: ν , ω_0 , and ω_p . Note that $\tau := \frac{1}{2\nu}$ and we use τ for convenience in the simulations, so the results reported below are in terms of τ .

For this time-domain parameter identification problem, the received data is observed electric field values at a discrete set of times. We collect these in a column vector V_{data} . Given a trial value of a vector of material parameters, say q , we may simulate a model of the system and collect electric field estimates at the same point in space and discrete times in order to form a vector $V_{model}(q)$. We again define the residual $R(q)$ and cost $F(q)$ as in (14) and (13). We intend to determine the dielectric parameter set q which minimizes the cost $F(q)$.

We use both Finkel's Direct global optimization program [6] and the MATLAB `lsqnonlin` function. Direct takes the n -dimensional rectangular region determined by given bounds and iteratively divides into smaller rectangles, checking for possible minimums. In this way, Direct is able to find the global minimum for functions with several local minima. On the other hand, `lsqnonlin` function uses gradient

methods to converge quickly to the nearest local minimum. Our strategy is to obtain an approximate solution using Direct to optimize ω_0^2 and ω_p , and then finish optimizing all three parameters with `lsqnonlin`.

For the first inverse problem, we consider how the deterministic model fits a distribution for single frequency signals. Data is synthesized from a model using a probability distribution with a range of 0.25 and then contaminated with normally-distributed random noise with a standard deviation of ζ .

We apply both deterministic and distributed fits for the 8×10^{15} frequency signal with noise of $\zeta = 2$ for comparison. Results are given in Table 4. Even though our method accurately recovered the true values of the material, the distributed fit was unable to significantly improve on the deterministic fit.

For the second inverse problem, we create data from the same distribution and attempt to apply deterministic and distributed fits. However, we now use a UWB as our interrogating signal:

$$f(t) = \sum_{i=1}^n \alpha_i \sin(f_i t) \tag{20}$$

where f_i are angular frequencies linearly spaced from 1×10^{14} to 1×10^{16} and α_i are weights determined by the beta distribution $\beta(1, 3)$. Results are given in Table 5 using $n = 100$. It is clearly harder for the deterministic model to fit a UWB than a single frequency signal.

The data supports the suggestion above that the distributed model struggles with estimating τ from the data. This is expected since a large change in τ corresponds to a small change in the cost function. Also, the distributed fit did make an appreciable difference over the cost of the deterministic fit. This agrees with our simulations in the frequency domain where the deterministic model was unable to fit parameters to the distributed permittivity over a spectrum of frequencies.

Table 4 Fit comparison: Freq = 8×10^{15} and $\zeta = 2$

Source	$\tau (1 \times 10^{-16})$	$\omega_0 (1 \times 10^{16})$	Range	$\omega_p (1 \times 10^{16})$	Cost	Norm. cost
Data	7	1.8	0.25	2	–	–
Det. fit	6.9489	1.7543	–	1.9697	23971	3.995
Dist. fit	6.9819	1.8049	0.2438	1.9984	23591	3.932

Table 5 Fit comparison: UWB with $\zeta = 2$

Source	$\tau (1 \times 10^{-16})$	$\omega_0 (1 \times 10^{16})$	Range	$\omega_p (1 \times 10^{16})$	Cost	Norm. cost
Data	7	1.8	0.25	2	–	–
Det. fit	6.4433	1.7757	–	2.0135	25825	4.304
Dist. fit	7.0650	1.7999	0.2493	1.9998	23441	3.907

6 Conclusion

We showed in the frequency domain that applying a distribution to ω_0^2 can produce significantly better fits of parameters to real data than the deterministic Lorentz model. In the time domain, we used Polynomial Chaos and finite differences with the first order Yee Scheme to discretize the Maxwell-random Lorentz system. In [1] it was shown that the Polynomial Chaos method converged quickly for the number of polynomials used in the expansion. For the inverse problem, we compared a single frequency interrogating signal with a UWB pulse. The distributed model only fit better than the deterministic model over a range of frequencies as implied by the complex permittivity plots in the frequency domain.

Acknowledgements This work was done during the 2017 Research Experiences for Undergraduates program in mathematics at Oregon State University, with support by National Science Foundation Grant DMS-1359173.

References

1. Alvarez, J., Fisher, A.: Approximating dispersive materials with parameter distributions in the Lorentz model. In: Gibson, N.L. (Ed.) REU Program at Oregon State University Proceedings (2017)
2. Banks, H.T., Gibson, N.L.: Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters. *Q. Appl. Math.* **64**(4), 749 (2006)
3. Banks, H.T., Kunisch, K.: Estimation Techniques for Distributed Parameter Systems. Springer Science & Business Media (1989)
4. Banks, H.T., Buksas, M.W., Lin, T.: Electromagnetic Material Interrogation Using Conductive Interfaces and Acoustic Wavefronts. SIAM (2000)
5. Banks, H.T., Catenacci, J., Hu, S.: Method comparison for estimation of distributed parameters in permittivity models using reflectance. *Eurasian J. Math. Comput. Appl.* **3**(2), 4–24 (2015)
6. Finkel, D.E.: DIRECT optimization algorithm user guide. Center for Research in Scientific Computation, North Carolina State University, 2 (2003)
7. Fisher, A., Alvarez, J., Gibson, N.L.: Analysis of methods for the Maxwell-random Lorentz model. *Results Appl. Math.* (2020). <https://doi.org/10.1016/j.rinam.2020.100098>
8. Gibson, N.L.: A polynomial chaos method for dispersive electromagnetics. *Commun. Comput. Phys.* **18**(5), 1234–1263 (2015)
9. Oughstun, K.E., Cartwright, N.A.: On the Lorentz-Lorenz formula and the Lorentz model of dielectric dispersion. *Opt. Express* **11**(13), 1541–1546 (2003)
10. Querry, M.R., Waring, R.C., Holland, W.E., Hale, G.M., Nijm, W.: Optical constants in the infrared for aqueous solutions of NaCl. *J. Opt. Soc. Am.* **62**(7), 849–855 (1972)
11. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press (2010)
12. Yee, K.: Numerical solution of initial boundary value problems involving Maxwell’s equations in isotropic media. *IEEE Trans. Antennas Propag.* **14**(3), 302–307 (1966)

Coulomb Explosion Imaging: Super-Resolution by Optical Properties of Electrostatics Lenses



David Babalola and C. Sean Bohun

Abstract Velocity-map imaging (VMI) is a popular technique in a Coulomb-explosion imaging experiment with the capacity to focus photo-fragments based on their initial velocity vectors. The VMI is capable of achieving this feat as a result of the system of electrostatic lenses with varying potential, which the photo-fragments have to transit. However, despite the focusing capability of the VMI, the measured time-of-flights of the photo-fragments still suffer from a temporal spread, which is a consequence of the initial velocity and spatial spread at the point of formation. To be able to improve the spatial-temporal resolution of the photo-fragments at the point of formation, there is a need for a better understanding of how the system of electrostatic lenses alter the trajectories of the photo-fragments between formation and detection to achieve a velocity map. Also, an expression is derived to resolve the spatial spread of the photo-fragment products.

Keywords Resolution · Velocity-map · Time-of-flights

1 Introduction

Coulomb-explosion imaging (CEI) is a technique used to determine a molecular structure by retrieving in coincidence the momenta of all the fragment ions from the parent molecule following the stripping of its valence electrons by an intense laser pulse [6–8]. In a CEI experiment, velocity-map imaging (VMI) is a technique that maps photo-fragments to the detector based on their initial velocities [5]. This simply means that, fragments with the same initial velocity will end up at the same spot on the detector, regardless of their initial positions at formation. In practice, the time of arrival of these positive ions are noted along with the positions of impacts on the

D. Babalola (✉) · C. S. Bohun
Ontario Tech University, 2000 Simcoe Street North, Oshawa, Canada
e-mail: david.babalola@ontariotechu.ca

C. S. Bohun
e-mail: sean.bohun@ontariotechu.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_33

361

detector. This information is vital for CEI, in particular for molecular reconstruction. From this information, correlation among the photo-fragments are determined to know those ions that are in coincidence i.e. the ions that undergo the same fragmentation process. The primary data used for the correlation of the fragments is the time of flight (TOF) [1, 6–8, 14], and the spatial data is used in the calculation of kinetic energy and angular distributions [5]. Hence, the need for a highly resolved measured data. According to [12, 13], some independent factors that impact the resolution of the time of flight are: responsiveness of the detection system, initial velocity spread, initial position spread and the field inhomogeneity. Research efforts have been directed at resolving some of these issues but some resolution concern still persists [2–5, 10, 11].

The central theme of this article is to look at how to improve resolution by adjusting the measured spatial-temporal information. This approach entails understanding how the electrostatic lens system creates velocity map as the ions transits the spectrometer, examine if the focusing of the fragments on the detector is mass-per-charge dependent and derive an expression for adjusting the spatial-temporal information. At the moment, this study does not consider the effects of open apertures on VMI lenses on the electric field, therefore, the potential is assumed linear between segments. The remainder of the article uses a paraxial approximation of the potential, which indirectly describes the fragment motion along axial direction. This is followed by the exploitation of the optical properties of the electrostatic lens in understanding time spread, an explanation of the simulation results, and a conclusion.

2 Paraxial Approximation

To improve spatial-temporal resolution in CEI experiments, there is a need for a better understanding of the dynamics of a fragment in a time-of-flight spectrometer (TOFS). Hence, we zoom-in on the influence of the electrostatics lenses on its trajectory under paraxial approximation. First of all, the electric field $\mathbf{E} = (E_r, E_\theta, E_z)$ is given by $-\nabla\phi$ and it satisfies the Laplace's equation $\nabla^2\phi = 0$ due to the lack of internal charge. In a paraxial approximation, the trajectory is assumed to be near the optical axis and we will show that indeed $E_r \ll E_z$ (i.e. the axial field component dominates the radial field component). The performance of a TOFS depends only on the ratios of the geometrical dimensions and potentials, which means any TOFS is scalable [13].

Let the radial and axial coordinates be scaled according to the physical dimension of the spectrometer of length L and radius R . In addition, the potential is scaled by the characteristic operating voltage V_R of the lens system so that

$$r = R\tilde{r}, \quad z = L\tilde{z}, \quad \phi = V_R\tilde{\phi}. \quad (1a)$$

The nondimensional parameter

$$\varepsilon = \frac{R}{L} \ll 1 \quad (1b)$$

for typical spectrometer. Using this parameter, and assuming an axi-symmetric, steady state, electric potential $\phi = \phi(r, Z)$ satisfies

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \phi}{\partial r} \right) + \varepsilon^2 \frac{\partial^2 \phi}{\partial z^2} = 0, \quad \phi(0, z) = 0 \quad (2)$$

where the tildes have been dropped for convenience. Expanding ϕ in a power series in ε , we let

$$\phi(r, z) = \sum_{j=0}^{\infty} \varepsilon^{2j} \phi_j(r, z). \quad (3)$$

Substituting this representation into expression (2), we have the following recursive equations,

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \phi_j}{\partial r} \right) = -\frac{\partial^2 \phi_{j-1}}{\partial z^2}, \quad \phi_j(0, z) = 0, \quad j = 0, 1, \dots \quad (4)$$

where for $\phi_{-1} \equiv 0$. Solving (4) for $j = 0$, we have

$$\phi_0(r, z) = c_0(z) \ln r + c_1(z). \quad (5a)$$

To ensure that the potential remains bounded along $r = 0$ gives

$$\phi_0(r, z) = c_1(z) = \phi(0, z) = \Phi(z). \quad (5b)$$

From this first result, the remaining terms can be solved explicitly and collecting these results,

$$\phi(r, z) = \sum_{j=0}^{\infty} \varepsilon^{2j} \frac{(-1)^j}{2^{2j}} \frac{r^{2j}}{(j!)^2} \frac{d^{2j} \Phi}{dz^{2j}}. \quad (5c)$$

The corresponding electric field, with the assumption of azimuthal symmetry,

$$\mathbf{E}(r, z) = -\nabla \phi(r, z) = -\left\langle \frac{\partial \phi}{\partial r}, \frac{1}{r} \frac{\partial \phi}{\partial \theta}, \frac{\partial \phi}{\partial z} \right\rangle = \langle E_r, E_\theta, E_z \rangle, \quad (6)$$

allows one to identify $E_\theta = 0$ and

$$E_r(r, z) = \frac{r}{2} \frac{d^2 \Phi}{dz^2} \varepsilon^2 + \mathcal{O}(\varepsilon^4), \quad E_z(0, z) = -\frac{d\Phi}{dz} + \frac{r^2}{4} \frac{d^3 \Phi}{dz^3} \varepsilon^2 + \mathcal{O}(\varepsilon^4). \quad (7)$$

3 Optical Property and Time Dispersion

Following the asymptotic expansion of the electric potential, there is a need to understand how the electrostatic lenses perturb the flight path to be able to achieve a velocity map. This exploration is carried out under paraxial approximation and the time dispersion is also analysed. Under the *strict* paraxial assumption for the potential, ($\varepsilon = 0$),

$$\mathbf{E} = -\nabla\phi = -\frac{d\phi}{dz}\hat{\mathbf{k}} \quad (8)$$

so that the changes in speed are restricted to the $\hat{\mathbf{k}}$ direction. Focusing on the corresponding $z = z(t)$ location, the trajectory of a fragment with mass m and charge q satisfies,

$$\frac{d^2z}{dt^2} = \frac{1}{2} \frac{d}{dz} \left(\frac{dz}{dt} \right)^2 = -\frac{q}{m} \frac{d\Phi}{dz}, \quad 0 \leq z \leq \ell \quad (9)$$

where $\ell > 0$ denotes the downstream edge of the lens assembly. Integrating from z_1 to $\ell \geq z_2 > z_1$, and letting $v = dz/dt$,

$$v^2(z_2) - v^2(z_1) = -\frac{2q}{m} (\Phi(z_2) - \Phi(z_1)). \quad (10)$$

With this expression we consider the effect of a dispersion in both the initial position and initial velocity. In particular, $v_0(t)$ with $z(t=0) = 0$, $v_0(z(t=0)) = 0$, and $v_1(t)$ with $z(t=0) = \Delta z$, $v_1(z(t=0)) = \Delta v$. At a position z , with $\Delta z < z \leq \ell$,

$$v_0^2(z) = -\frac{2q}{m} (\Phi(z) - \Phi(0)), \quad (11a)$$

$$v_1^2(z) = -\frac{2q}{m} (\Phi(z) - \Phi(\Delta z)) + (\Delta v)^2 = v_0^2(z) - v_0^2(\Delta z) + (\Delta v)^2, \quad (11b)$$

where we have assumed that $d\phi/dz < 0$ for the entire interval. The time to reach $z = \ell$ in either scenario is therefore

$$T_0 = \int_0^\ell \frac{dt}{v_0(z)}, \quad T_1 = \int_{\Delta z}^\ell \frac{dt}{v_1(z)}, \quad (12a)$$

and consequently,

$$\Delta T = T_1 - T_0 = \int_{\Delta z}^\ell \frac{dz}{(v_0^2(z) - v_0^2(\Delta z) + (\Delta v)^2)^{1/2}} - \int_0^\ell \frac{dz}{v_0(z)}. \quad (12b)$$

It is convenient to define an auxiliary displacement Δz_1 to be that position where v_1 vanishes so that $v_1(\Delta z_1) = 0$. According to (11b),

$$v_0^2(\Delta z_1) - v_0^2(\Delta z) + (\Delta v)^2 = 0. \tag{13}$$

With this definition we can rewrite (12b) as

$$\begin{aligned} \Delta T = & - \int_0^\ell \frac{dz}{v_0(z)} + \int_{\Delta z_1}^\ell \frac{dz}{(v_0^2(z) - v_0^2(\Delta z_1))^{1/2}} \\ & - \text{sgn}(\Delta v) \int_{\Delta z_1}^{\Delta z} \frac{dz}{(v_0^2(z) - v_0^2(\Delta z_1))^{1/2}} \end{aligned} \tag{14}$$

where the $\text{sgn}(\Delta v)$ factor compensates for the ordering of Δz and Δz_1 .

Each of these integrals are now approximated assuming Δz and Δz_1 are close to zero. In particular,

$$(v_0^2(z) - v_0^2(\Delta z_1))^{-1/2} = (2a_0(z - \Delta z_1))^{-1/2} \left(1 - \frac{b_0}{8a_0}(z + \Delta z_1) + \mathcal{O}(z^2) \right) \tag{15a}$$

where

$$a_0 = \frac{1}{2} \left. \frac{dv_0^2(z)}{dz} \right|_{z=0}, \quad b_0 = \left. \frac{d^2v_0^2(z)}{dz^2} \right|_{z=0}. \tag{15b}$$

Integrating over $[\Delta z_1, \Delta z]$ we find

$$\begin{aligned} \text{sgn}(\Delta v) \int_{\Delta z_1}^{\Delta z} \frac{dz}{(v_0^2(z) - v_0^2(\Delta z_1))^{1/2}} &= \text{sgn}(\Delta v) \left(\frac{2(\Delta z - \Delta z_1)}{a_0} \right)^{1/2} (1 + \Lambda(\Delta z)) \\ &= \frac{\Delta v}{a_0} (1 + \Lambda(\Delta z)), \quad \Lambda(\Delta z) = \frac{b_0(\Delta z + 5\Delta z_1)}{24a_0} + \mathcal{O}((\Delta z)^2), \end{aligned} \tag{16}$$

where expression (13) expanded about $z = 0$ gives

$$\Delta z_1 = \Delta z - \frac{(\Delta v)^2}{2a_0} + \mathcal{O}((\Delta z)^2). \tag{17}$$

Similarly, detailed in the Appendix,

$$\int_0^\ell \frac{dz}{v_0(z)} - \int_{\Delta z_1}^\ell \frac{dz}{(v_0^2(z) - v_0^2(\Delta z_1))^{1/2}} = \Delta z \left(\frac{1}{v_0(\ell)} + \int_0^\ell \frac{a(z) - a_0}{v_0^3(z)} dz \right) + \mathcal{O}((\Delta z)^2). \tag{18}$$

Combining (16) and (18) the total time dispersion is to first order in Δz and Δv ,

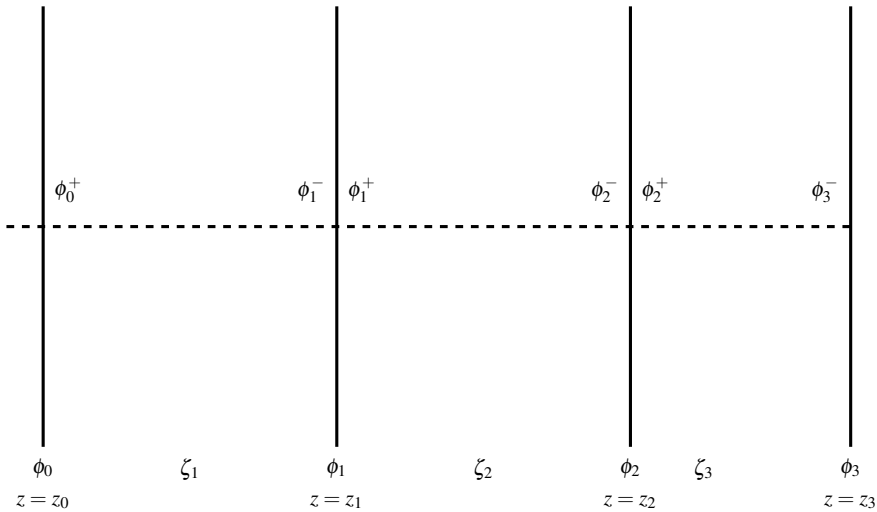


Fig. 1 A skeleton of a system of $N = 3$ electrostatic lenses with multiple segments $\{\zeta_k\}_{k=1}^3$ showing how the potentials immediately before and after an electrode. Within each segment, the potential is linear and the acceleration is constant

$$\Delta T = \frac{\Delta v}{a_0} + \Delta z \left(\frac{1}{v_0(\ell)} + \int_0^\ell \frac{a(z) - a_0}{v_0^3(z)} dz \right), \quad a(z) = \frac{1}{2} \frac{dv_0^2(z)}{dz}. \quad (19)$$

In conclusion, this analysis shows that the dispersion due to Δv cannot be eliminated but is minimized by increasing the initial acceleration. In contrast the dispersion due to the spread in initial location, Δz , can be eliminated by choosing potential so that the condition

$$\int_0^\ell \frac{a_0 - a(z)}{v_0^3(z)} dz = \frac{1}{v_0(\ell)} \quad (20)$$

is satisfied.

To understand how a series of lenses impact the trajectory of a fragment between formation and detection, we consider an N -segment TOFS. To this end, consider a collection of N segments $\zeta_k = [z_{k-1}, z_k], k = 1, 2, \dots, N$. with a specified velocity, $v(z_k) = v_k$. Figure 1 illustrates a TOFS for $N = 3$. Assuming a constant acceleration a_k for segment ζ_k then

$$a_k = \frac{v_k^2 - v_{k-1}^2}{2(z_k - z_{k-1})}, \quad t_k = \frac{2(z_k - z_{k-1})}{v_k + v_{k-1}}, \quad T_0 = \sum_{k=1}^N t_k \quad (21)$$

where t_k is the time to traverse segment ζ_k and T_0 is the time taken to transit all of the segments.

The axial motion is governed by (9) for each segment. Focusing on the radial motion we have

$$\frac{d^2 r}{dt^2} = \frac{q}{2m} \frac{d^2 \Phi}{dz^2} r \quad (22)$$

where the constant axial acceleration, $d\Phi/dz = \text{const}$, translates into a constant velocity in the radial direction. Considering a fragment that enters ζ_k at a radius of r_{k-1} , and velocity $dr/dt (z = z_{k-1}) = \dot{r}_{k-1}$ will exit at radius $r_k = r_{k-1} + \dot{r}_{k-1} t_k$. As a transition matrix,

$$\begin{pmatrix} r_k \\ \dot{r}_k \end{pmatrix} = \begin{pmatrix} 1 & t_k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_{k-1} \\ \dot{r}_{k-1} \end{pmatrix} = P_k \begin{pmatrix} r_{k-1} \\ \dot{r}_{k-1} \end{pmatrix}, \quad (23)$$

for the segment ζ_k . As the fragment crosses from one segment to the next, there is jump in the potential which updates the velocity and using (22) gives the jump condition

$$\dot{r}_k^+ - \dot{r}_k^- = \frac{q}{2m} \frac{r_k^-}{v_k} (\phi_k^+ - \phi_k^-), \quad r_k^+ = r_k^-, \quad (24)$$

or as a matrix we have

$$\begin{pmatrix} r_k \\ \dot{r}_k \end{pmatrix}^+ = \begin{pmatrix} 1 & 0 \\ \frac{q}{2m} (\phi_k^+ - \phi_k^-) & 1 \end{pmatrix} \begin{pmatrix} r_k \\ \dot{r}_k \end{pmatrix}^- = L_k \begin{pmatrix} r_k \\ \dot{r}_k \end{pmatrix}^-. \quad (25)$$

Combining (23) and (25) for the 3-lens TOFS system,

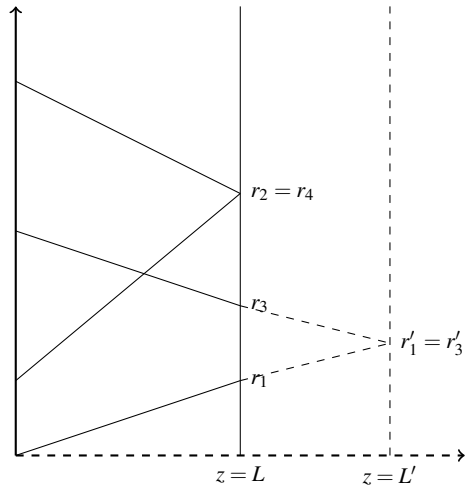
$$\begin{pmatrix} r_3 \\ \dot{r}_3 \end{pmatrix} = M \begin{pmatrix} r_0 \\ \dot{r}_0 \end{pmatrix}, \quad M = P_3 L_2 P_2 L_1 P_1 = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \quad (26)$$

is the transformation matrix connecting the initial radial position and velocity of a fragment with its final radial position and velocity at the detector. Since P_k and L_k have unit determinant, the matrix M itself has unit determinant so that the transformation is uniquely invertible.

If $M_{12} = 0$ then fragments leaving the same position on at their point of formation, will hit the same spot on the detector, irrespective of their initial velocity. That is, there is a one-to-one relationship between the radial position on the detector and its radial position at its starting point [9, 12, 13]. Similarly, if $M_{11} = 0$, and $M_{12} \neq 0$, the TOFS can achieve a velocity map in which fragments with the same initial velocity can be mapped to the same position, irrespective of their initial position. This second property is utilized in a VMI lens setup by placing a position-sensitive detector in the focal position. In this latter case

$$M_{11} = 1 + \frac{q}{2m} f(\{t_k\}_{k=1}^3, \{\phi_k^-\}_{k=1}^3, \{\phi_{k-1}^+\}_{k=1}^3) \quad (27)$$

Fig. 2 An exaggerated hit positions on a position-sensitive detector. All measurements are with reference to the center-line of the TOFS, i.e. the dashed horizontal line. For a detector placed at $z = L$, $r_2 = r_4$ is a hit position for two smaller masses with the same initial velocity, and r_1, r_3 are hit positions for another two larger masses with the same initial velocity. Adjustment is made to the positions r_1, r_3 so that they have a common hit position $r'_1 = r'_3$.



for a continuous function f . Therefore, placing the detector at a focal point for a fragment with a small value of m would not properly focus a much heavier fragment. So, there is a need to adjust the measured data for the fragments to account for this deficiency. This effect is shown in Fig. 2 where four fragments arrive at a detector with one pair, fragments 2 and 4 of say mass $m = m_1$, arriving at the focal $z = L$ and fragments 1 and 3 of mass $m = m_2 \neq m_1$, coinciding at the focal plane $z = L'$.

An example of this effect is simulated in Fig. 3 where the detector is re-positioned from $z = L$ to $z = L'$. In this simulation, the flight of four ions, two pairs, is calculated with a pair having the same mass per charge. For each pair of ion, one is back-scattered and the other is forward-scattered. Also, the position of each ion is rotated to account for all possible angles of flight. The measured spatial-temporal data for the fragments are re-calculated to compensate the fact that a single fixed detection plane would not provide an adequate resolution achievable by the detection system. The new final positions are calculated with $r' = r - \Delta r$, where $r' \in (r'_1, r'_3)$, $r \in (r_1, r_2, r_3, r_4)$, \bar{r} is a mean position and $\Delta r = r - \bar{r}$ which is the gap between the position of a fragment and the mean position of all of the fragments in the same ring of width $2\Delta r$. The spatial spread is improved as can be seen in the right frame of Fig. 3. This approach would reduce spread or dispersion that hitherto remains despite velocity mapping.

4 Conclusion

Using a paraxial approximation for a TOFS trajectories of system are shown to be tuneable, giving both the options for imaging that is invariant to the initial fragment velocity as well as velocity map imaging where the image does not depend on the

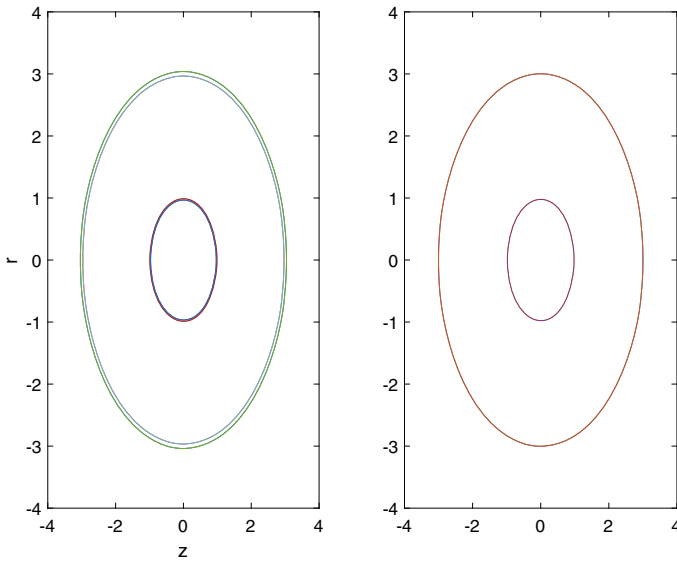


Fig. 3 An adjusted positions for fragments. The outer rings have a width that is attributable to a velocity spread in a VMI lens. After the position re-adjustment, the width is eliminated, hence, a better resolution

initial position. The time-of-flight is shown to have an inherent dispersion due to velocity variations which can be minimized with a large initial acceleration in the lens system, but never fully removed. Dispersion due to the initial fragment position can be completely eliminated. Further improvement in resolution is achievable through a re-positioning procedure that takes the mass of the various fragments into consideration. Also, a future improvement to the model that considers the potential as non-linear between segments will have a far-reaching applications to a true VMI instrument.

Appendix

The expression of the form,

$$I(s) = \int_s^L \frac{dx}{\sqrt{f(x) - f(s)}} - \int_0^L \frac{dx}{\sqrt{f(x) - f(0)}},$$

can be re-written with a change of variable $x' = x - s$ in the first integral and the domain of the second integral partitioned as $[0, L - s] \cup [L - s, L]$. The resulting representation is,

$$\int_0^{L-s} \left(\frac{dx}{\sqrt{f(x+s) - f(s)}} - \frac{dx}{\sqrt{f(x) - f(0)}} \right) - \int_{L-s}^L \frac{dx}{\sqrt{f(x) - f(0)}}.$$

The first integrand in the above expression is expanded in power series in s as,

$$I_1(s) = -\frac{s}{2} \frac{f'(x) - f'(0)}{\sqrt{(f(x) - f(0))^3}} + \mathcal{O}(s^2), \quad s \ll 1,$$

(the assumption $s \ll 1$ is sufficient for our application, see (18) in the text) and the second integral as,

$$I_2(s) = \int_{L-s}^L \frac{dx}{\sqrt{f(x) - f(0)}} = \int_0^s \frac{dy}{\sqrt{f(L+y-s) - f(0)}},$$

with the substitution $y = x - L + s$. Under the assumption that, $s \ll L$,

$$I_1(s) = -\frac{1}{2} \frac{f'(x) - f'(0)}{(f(x) - f(0))^{3/2}} s + \mathcal{O}(s^2), \quad I_2(s) = \frac{s}{\sqrt{f(L) - f(0)}} + \mathcal{O}(s^2).$$

Therefore, the integral $I(s)$ is evaluated to the first order as,

$$I(s) = -s \left(\frac{1}{2} \frac{f'(x) - f'(0)}{(f(x) - f(0))^{3/2}} + \frac{1}{\sqrt{f(L) - f(0)}} \right) + \mathcal{O}(s^2).$$

References

1. Bocharova, I.A., Alnaser, A.S., Thumm, U., Niederhausen, T., Ray, D., Cocke, C.L., Litvinyuk, I.V.: Time-resolved Coulomb-explosion imaging of nuclear wave-packet dynamics induced in diatomic molecules by intense few-cycle laser pulses. *Phys. Rev. A* **83**, 013417 (2011)
2. Brouard, M., Campbell, E.K., Johnsen, A.J., Vallance, C., Yuen, W.H., Nomerotski, A.: Velocity map imaging in time of flight mass spectrometry. *Rev. Sci. Instrum.* **79**, 123115 (2008)
3. Christensen, L., Christiansen, L., Shepperson, B., Stapelfeldt, H.: Deconvoluting nonaxial recoil in Coulomb explosion measurements of molecular axis alignment. *Phys. Rev. A* **94**, 023410 (2016)
4. Coles, J.N., Guilhaus, M.: Resolution limitations from detector pulse width and jitter in a linear orthogonal-acceleration time-of-flight mass spectrometer. *J. Am. Soc. Mass Spectrom.* **5**, 772–778 (1994)
5. Eppink, A.T.J.B., Parker, D.H.: Velocity map imaging of ions and electrons using electrostatic lenses: application in photoelectron and photofragment ion imaging of molecular oxygen. *Rev. Sci. Instrum.* **68**, 3477–3484 (1997)
6. Frasiniski, L.J., Codling, K., Hatherly, P.A.: Covariance mapping: a correlation method applied to multiphoton multiple ionization. *Science* **246**, 1029–1031 (1989)
7. Frasiniski, L.J.: Covariance mapping techniques. *J. Phys. B: At. Mol. Opt. Phys.* **49**, 152004 (2016)
8. Gagnon, J., Lee, K.F., Rayner, D.M., Corkum, P.B., Bhardwaj, V.R.: Coincidence imaging of polyatomic molecules via laser-induced Coulomb explosion. *J. Phys. B: At. Mol. Opt. Phys.* **41**, 215104 (2008)

9. Gerrard, A., Burch, J.M.: *Introduction to Matrix Methods in Optics*. Courier Corporation (1994)
10. Guilhaus, M.: Special feature: tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts. *J. Mass Spectrom.* **30**, 1519–1532 (1995)
11. Karas, M.: 'Time-of-flight mass spectrometer with improved resolution,' W. C. Wiley and I. H. McLaren, *Rev. Sci. Instrum.*, 26, 1150 (1955). *J. Mass Spectrom.* **32**, 1–11 (1997)
12. Meron, M.: Design and optimization of time-of-flight spectrometers. *Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip.* **291**, 637–645 (1990)
13. Meron, M.: Design and optimization of time-of-flight spectrometers part II. *Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrom. Detect. Assoc. Equip.* **330**, 259–267 (1993)
14. Slater, C.S., Blake, S., Brouard, M., Lauer, A., Vallance, C., Bohun, C.S., Christensen, L., Nielsen, J.H., Johansson, M.P., Stapelfeldt, H.: Coulomb-explosion imaging using a pixel-imaging mass-spectrometry camera. *Phys. Rev. A* **91**, 053424 (2015)

Error Correction for Correlated Quantum Systems



Mark Byrd, Alvin Gonzales, Daniel Dilley, and Purva Thakre

Abstract Modeling open quantum systems is a difficult task for many experiments. A standard method for modeling open system evolution uses an environment that is initially uncorrelated with the system in question, evolves the two unitarily, and then traces over the bath degrees of freedom to find an effective evolution of the system. This model can be insufficient for physical systems that have initial correlations. Specifically, there are evolutions $\rho_S = \text{tr}_E(\rho_{SE}) \rightarrow \rho'_S = \text{tr}_E(U\rho_{SE}U^\dagger)$ which cannot be modeled as $\rho_S = \text{tr}_E(\rho_{SE}) \rightarrow \rho'_S = \text{tr}_E(U\rho_S \otimes \rho_E U^\dagger)$. An example of this is $\rho_{SE} = |\Phi^+\rangle\langle\Phi^+|$ and $U_{SE} = \text{CNOT}$ with control on the environment. Unfortunately, there is no known method of modeling an open quantum system which is completely general. We first present some restrictions on the availability of completely positive (CP) maps via the standard prescription. We then discuss some implications a more general treatment would have for quantum control methods. In particular, we provide a theorem that restricts the reversibility of a map that is not completely positive (NCP). Let Φ be NCP and $\tilde{\Phi}$ be the corresponding CP map given by taking the absolute value of the coefficients in Φ . The theorem shows that the CP reversibility conditions for $\tilde{\Phi}$ do not provide reversibility conditions for Φ unless Φ is positive on the domain of the code space.

Keywords Quantum error correction · Quantum control · Open quantum systems

1 Introduction

Precise modeling and control of quantum systems will be required for quantum technologies. This includes quantum computers, quantum cryptography, and quantum simulation of quantum systems. Unfortunately, even though great progress has been

M. Byrd (✉) · A. Gonzales · D. Dilley · P. Thakre
Southern Illinois University Carbondale, 1263 Lincoln Drive, Carbondale, IL 62901-6899, USA
e-mail: mbyrd@siu.edu

A. Gonzales
e-mail: agonza@siu.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_34

made in this area, there are still questions concerning the types of possible evolutions and how to describe them. In the case that the system and environment have no prior correlations, there is a standard prescription for describing the evolution. Let the system (environment) state be ρ_S (ρ_E). Supposing that the two evolve under a joint unitary transformation U_{SE} , then a map can be defined by

$$\Phi(\rho_S) = \text{tr}_E(U_{SE}\rho_S \otimes \rho_E U_{SE}^\dagger). \quad (1)$$

This map is not only positive (It maps all positive operators to positive operators.), it is also completely positive (The extended map $\mathbb{I}_n \otimes \Phi(\rho)$ is positive for all n and any positive input ρ .) This also provides a way to model a quantum system.

However, not all evolutions of a quantum system are able to be described this way. In particular, the assumption of an initial product state may not be satisfied. In this case, some discussions have arisen in the literature about what one should do if the standard assumption of an initially uncorrelated state no longer applies [1–13]. This is very relevant given that such examples are not difficult to find [14].

In general, the evolution of a system can be defined by a dynamical map, A , where we first vectorize the system density matrix [15]. The vectorization is done by writing all the elements as a column vector. For a single qubit state, this is given by

$$\text{vec}(\rho) = \begin{bmatrix} \rho_{00} \\ \rho_{01} \\ \rho_{10} \\ \rho_{11} \end{bmatrix}. \quad (2)$$

The transformation is then done on this vectorized form and is given by

$$\rho' = A\rho. \quad (3)$$

Using the restrictions for a valid density matrix, it being Hermitian, positive semi-definite and having trace one, the restrictions on the A matrix are given by

$$A_{rs,r's'} = (A_{s'r',r's})^*, \quad (4)$$

$$\sum_{rsr's'} x_r^* x_s A_{rs,r's'} y_{r'} y_{s'}^* \geq 0, \quad (5)$$

and

$$\sum_r A_{rr,r's'} = \delta_{r's'}, \quad (6)$$

respectively.

These conditions can be translated to an equivalent B matrix by just relabeling. Let

$$B_{rr',ss'} \equiv A_{rs,r's'} \tag{7}$$

Then, the conditions are hermiticity

$$B_{rr',ss'} = (B_{ss',rr'})^*, \tag{8}$$

positivity

$$\sum_{rsr's'} z_{rr'}^* B_{rr',ss'} z_{ss'} \geq 0, \tag{9}$$

and trace preserving

$$\sum_r B_{rr',rs'} = \delta_{r's'} \tag{10}$$

Since B is Hermitian, it has an eigenvector/eigenvalue decomposition

$$B_{r'r,s's'} \rho_{rs} = \sum_{\alpha} \gamma(\alpha) C_{r'r}^{\alpha} \rho_{rs} (C_{s's}^{\alpha})^*,$$

where the C are the eigenvectors and γ the eigenvalues.

One may also write this as

$$\Phi(\rho) = B\rho = \sum_{\alpha} \eta_{\alpha} A_{\alpha} \rho A_{\alpha}^{\dagger} \left(= \sum_{\alpha} A_{\alpha} \rho A_{\alpha}^{\dagger}, \forall \eta_{\alpha} = 1 \right), \tag{11}$$

where $A_{\alpha} \equiv \sqrt{|\gamma|} C^{\alpha}$ so that $\eta_{\alpha} = \pm 1$. It is known that the map is completely positive (CP) if and only if all $\eta_{\alpha} = 1$.

This form is often called the ‘‘Operator-Sum representation’’, or ‘‘Kraus decomposition’’ and is often used to describe open-system quantum dynamics.

2 Freedom in the Operator-Sum Representation

It is important to realize that the operator-sum decomposition, Eq. (11), is not unique and this non-uniqueness can be useful for finding different operator bases. This freedom is often called the ‘‘unitary freedom’’ [16].

Unitary Theorem: The form of a completely positive Hermiticity-preserving map, $\Phi(\rho) = \sum_{\alpha} A_{\alpha} \rho A_{\alpha}^{\dagger}$, defined by operators $\{A_{\alpha}\}$ is not unique, but the operators $\{F_{\beta}\}$

give the same map, if and only if there is a unitary matrix with elements $u_{\alpha\beta}$ such that $F_\beta = \sum_\alpha u_{\beta\alpha} A_\alpha$, $\forall \beta$.

This theorem can be used to prove the error-correcting code conditions below.

The more general map, the map that may be not completely positive (NCP), has a freedom in it as well. This is called the “pseudo-unitary freedom” [17].

Pseudo-Unitary Theorem: The form of a Hermiticity-preserving map,

$$\Phi(\rho) = \sum_\alpha \eta_\alpha A_\alpha \rho A_\alpha^\dagger,$$

defined by $\{A_\alpha\}$ and $\{\eta_\alpha\}$ is not unique, but the operators $\{F_\beta\}$ give the same map, if and only if there is a pseudo-unitary matrix with elements $u_{\alpha\beta}$ such that $F_\alpha = \sum_\beta u_{\alpha\beta} A_\beta$, $\forall \alpha$. The signature of the matrix $(u_{\alpha\beta}) \in U(p, q)$ is determined by the number of input and output elements in the sets $\{A_\alpha\}$ and $\{F_\beta\}$.

Note that a unitary matrix, V , is defined by the equation $VIV^\dagger = I$, whereas a pseudounitary matrix, U , is defined by the equation $U\eta U^\dagger = \eta$. In general, there are many choices for η . However, in our case, $\eta = \text{diag}(1, 1, \dots, 1, -1, -1, \dots, -1)$ where there are p ones and q minus ones.

3 Modeling Open Quantum Systems

The standard prescription, Eq. (1), is used to justify completely positive maps, and often suffices for modeling quantum systems. However, it is clear that it is not the most general possible evolution. Many people, including the recent work of Pechukas, which spurred much discussion [1], have pointed out that a more general evolution may be derived from a potentially correlated system and environment:

$$\Phi(\rho_S) = \text{tr}_E(U_{SE}\rho_{SE}U_{SE}^\dagger). \tag{12}$$

Finding examples which do not obey the assumption of an uncorrelated system and environment is not difficult. Consider the following two qubit example. Suppose that initial and final states of the system are known to be, respectively,

$$\rho_S = (1/2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\rho'_S = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Further assume that it is known that they evolve according to a system-environment coupling

$$U_{SE} = (1/\sqrt{2}) \begin{pmatrix} -i & 0 & 0 & -i \\ 0 & -i & -i & 0 \\ -i & 0 & 0 & i \\ 0 & -i & i & 0 \end{pmatrix}.$$

Then it is easy to show that there is no state ρ_E such that Eq. (1) is satisfied. This example can be shown to be robust to initial condition variations, as well as variations in the unitary transformation. This makes it experimentally verifiable.

Furthermore, finding such examples is not difficult. Consider a transformation from

$$\rho_S = \text{Tr}_E(\rho_{SE})$$

to

$$\rho'_S = \text{Tr}_B(U_{SE}\rho_{SE}U_{SE}^\dagger).$$

We say this is *U-generated* by U_{SE} . The set of local unitary transformations will be denoted LU, the set of unitaries that are equivalent via local unitaries to the swap unitary will be denoted as SWAP, and the set of unitary transformations that are equivalent via local unitaries to a controlled unitary will be denoted U_{C2} . Then we have the following theorem [14].

Theorem: Suppose that the system and environment consist of two qubits. Every U -generated physical transformation $\rho_S \rightarrow \rho'_S$ can be U -generated by a product state iff U belongs to $\text{LU} \cup \text{SWAP}$. If U belongs to U_{C2} , the transformation can be U -generated by a quantum-classical state. On the other hand, if U does not belong to $\text{LU} \cup \text{SWAP} \cup U_{C2}$, then there exist physical transformations that cannot be U -generated by any initial separable state.

Therefore, there are plenty of examples where the standard prescription fails. This is our motivation for studying evolutions that do not necessarily correspond to a completely positive map.

4 Reversing a Quantum Operation Corresponding to a Completely Positive Map

The reversibility of a quantum operation depends on the operation elements satisfying certain conditions. These conditions are known as the *quantum error correcting code conditions*. There are several ways to state these conditions, one is to consider a map of the form of Eq. (11) when the map is completely positive with operation elements A_α , and some logical (encoded states) $|i_L\rangle, |j_L\rangle$ [18]

$$\langle i_L | A_\alpha^\dagger A_\beta | j_L \rangle = m_{\alpha\beta} \delta_{ij}.$$

This has an intuitive interpretation as a “disjointness condition.” It states that one state $|i_L\rangle$ acted on by one operator A_α cannot have any overlap with another state $|j_L\rangle$ acted on by another error A_β .

One can show that this condition is equivalent to the following condition [19]

$$PA_\alpha^\dagger A_\beta P = c_{\alpha\beta} P,$$

where P is a projector onto the code space and c is a Hermitian matrix.

5 Reversing a Quantum Operation that Is Not a Completely Positive Map

We first show that it is possible to reverse a map that is NCP.

Example 1 Let the NCP map be the three qubit map

$$\Phi(\rho) = c_0\rho + c_1 \sum_i X_i \rho X_i - c_2 |010\rangle \langle 010| \rho |010\rangle \langle 010|, \tag{13}$$

where X_i is the Pauli matrix σ_x acting on qubit i and $c_0 + 3c_1 = 1$ and $0 < c_2 < 1$. Thus, Φ is a trace decreasing map. Suppose we know that Φ has occurred. Then, the projector onto the code space is

$$P = |000\rangle \langle 000| + |111\rangle \langle 111| \tag{14}$$

and the recovery map is

$$R(\rho) = P\rho P + \sum_i P X_i \rho X_i P. \tag{15}$$

It is easy to check that any state $P\rho P$ in the code space is recovered.

CP and NCP maps are closely related and in the paper by Shabani and Lidar [20], they state:

Corollary 1 Consider a Hermitian noise map

$$\Phi_H(\rho) = \sum_{i=1}^N \eta_i A_i \rho A_i^\dagger$$

and associate to it a CP map

$$\tilde{\Phi}_{CP}(\rho) = \sum_{i=1}^N |\eta_i| A_i \rho A_i^\dagger.$$

Then any quantum error correcting code and corresponding CP recovery map for $\tilde{\Phi}_{CP}(\rho)$ are also a quantum error correcting code and CP recovery map for $\Phi_H(\rho)$.

Their corollary gives a result which is proportional to the original density operator on average. However, the standard procedure for a quantum error correction, which reverses a quantum operation, proceeds in two steps. The first is to measure an error syndrome which identifies the error. The second step is the recovery operation. Since the first projects out one of the terms in the sum, the terms in the sum should all be positive if they are independent. Otherwise, they can give a negative result for the measurement, which corresponds to a negative probability for the result to occur. We deem this nonphysical.

Theorem 1 *Suppose, using the pseudo-unitary (PU) degree of freedom, that*

$$P F_i^\dagger F_j P = d_{ij} P$$

and

$$\Phi(\rho) = \Phi_1(\rho) - \Phi_2(\rho),$$

where $F_i = u_{ij} A_j$, $\{u_{ij}\} \in PU$, $\Phi_2(P\rho P) \neq 0$, and $\{d_{ij}\}$ is diagonal. Then $\Phi(P\rho P)$ is not positive, i.e., the code space is not in the domain of the error map.

Sketch of Proof: Let our input density matrix be $P\rho P$, i.e., in the code space. The proof relies on the orthogonality of the rotated code space. The code space projector P , when acted on by the individual operators F_i are rotated to a set of orthogonal projectors due to the error correcting condition. From the polar decomposition, we have

$$F_i P = U_i \sqrt{P F_i^\dagger F_i P} = \sqrt{d_{ii}} U_i P \tag{16}$$

This is actually a rotation on P . Thus, we can define

$$P_i \equiv U_i P U_i^\dagger \tag{17}$$

and when $i \neq j$ we get

$$P_j P_i = 0. \tag{18}$$

This means that we can pick out individual terms in the map.

Any NCP map can be written as the difference of two completely maps because we can group the negative terms and factor out the minus sign. Since the map $\Phi(\rho) = \Phi_1(\rho) - \Phi_2(\rho)$ and $\Phi_2(P\rho P) \neq 0$, we can get a measurement result P_i which corresponds to an outcome $P_i U_i \rho U_i^\dagger P_i$ by measuring the output density matrix in the $\{P_k\}$ basis. For $P_i U_i \rho U_i^\dagger P_i \in \Phi_2(P\rho P)$, this measurement probability is negative because the probability is given by

$$\text{tr}(-|d_{ii}\rangle P_i U_i \rho U_i^\dagger P_i) = -|d_{ii}|. \quad (19)$$

Since valid density matrices are positive semi-definite, the code space is not in the domain of $\Phi(\rho)$.

6 Discussion/Conclusions

The general problem of reversing the open-system evolution of a quantum system is an important open problem. Here we have provided a restriction on the ability to perform such an operation. In particular, we have shown that it is possible to arrive at a nonphysical result when attempting to use the same recovery operation for a map that is not completely positive as for the corresponding positive one. Furthermore, our theorem shows that there is a general restriction on the type of encoding that one may hope to use for reversing the quantum operations.

The general problem of how to reverse a quantum operation is still unsolved. However, we hope to present results elsewhere that can, in particular instances, enable the reversibility. Since the control of quantum systems is required for reliable quantum devices, we hope the results presented here, and in our future work, will help with the development of strategies for quantum control.

Acknowledgements Funding for this research was provided by the NSF, MPS under award number PHY-1820870.

References

1. Pechukas, P.: Phys. Rev. Lett. **73**, 1060 (1994). <https://doi.org/10.1103/PhysRevLett.73.1060>
2. Jordan, T.F., Shaji, A., Sudarshan, E.C.G.: Phys. Rev. A **70**, 052110 (2004). <https://doi.org/10.1103/PhysRevA.70.052110>
3. Rodríguez-Rosario, C.A., Modi, K., Meng Kuah, A., Shaji, A., Sudarshan, E.C.G.: J. Phys. A: Math. Theor. **41**(20), 205301 (2008). <https://doi.org/10.1088/1751-8113/41/20/205301>
4. Modi, K., Rodríguez-Rosario, C.A., Aspuru-Guzik, A.: Phys. Rev. A **86**, 064102 (2012). <https://doi.org/10.1103/PhysRevA.86.064102>
5. Liu, L., Tong, D.M.: Phys. Rev. A **90**, 012305 (2014). <https://doi.org/10.1103/PhysRevA.90.012305>
6. Buscemi, F.: Phys. Rev. Lett. **113**, 140502 (2014). <https://doi.org/10.1103/PhysRevLett.113.140502>
7. Dominy, J.M., Shabani, A., Lidar, D.A.: Quantum Inf. Process. **15**(1), 465 (2016). <https://doi.org/10.1007/s11128-015-1148-0>
8. Dominy, J.M., Lidar, D.A.: Quantum Inf. Process. **15**(4), 1349 (2016). <https://doi.org/10.1007/s11128-015-1228-1>
9. Allen, J.M.A., Barrett, J., Horsman, D.C., Lee, C.M., Spekkens, R.W.: Phys. Rev. X **7**, 031021 (2017). <https://doi.org/10.1103/PhysRevX.7.031021>
10. Schmid, D., Ried, K., Spekkens, R.W.: Phys. Rev. A **100**, 022112 (2019). <https://doi.org/10.1103/PhysRevA.100.022112>
11. Barrett, J., Lorenz, R., Oreshkov, O.: [arXiv:1906.10726](https://arxiv.org/abs/1906.10726) (2019)

12. Hartmann, R., Strunz, W.T.: Phys. Rev. A **101**, 012103 (2020). <https://doi.org/10.1103/PhysRevA.101.012103>
13. Barrett, J., Lorenz, R., Oreshkov, O.: [arXiv:2002.12157](https://arxiv.org/abs/2002.12157) (2020)
14. Chitambar, E., Abu-Nada, A., Ceballos, R., Byrd, M.: Phys. Rev. A **92**, 052110 (2015). <https://doi.org/10.1103/PhysRevA.92.052110>
15. Sudarshan, E.C.G., Mathews, P.M., Rau, J.: Phys. Rev. **121**, 920 (1961). <https://doi.org/10.1103/PhysRev.121.920>
16. Choi, M.D.: Linear Algebra Appl. **10**(3), 285 (1975). [https://doi.org/10.1016/0024-3795\(75\)90075-0](https://doi.org/10.1016/0024-3795(75)90075-0)
17. Ou, Y.C., Byrd, M.S.: Phys. Rev. A **82**, 022325 (2010). <https://doi.org/10.1103/PhysRevA.82.022325>
18. Knill, E., Laflamme, R.: Phys. Rev. A **55**, 900 (1997). <https://doi.org/10.1103/PhysRevA.55.900>
19. Nielsen, M.A., Caves, C.M., Schumacher, B., Barnum, H.: Proc. R. Soc. Lond. A **454**, 277 (1998). <https://doi.org/10.1098/rspa.1998.0160>
20. Shabani, A., Lidar, D.A.: Phys. Rev. A **80**, 012309 (2009). <https://doi.org/10.1103/PhysRevA.80.012309>

Numerical Investigation of VAWT Airfoil Shapes on Power Extraction and Self-starting Purposes



Sajad Maleki Dastjerdi, Amir HormoziNejad, Kobra Gharali,
and Jatin Nathwani

Abstract The effects of airfoil shapes on the power coefficient and the torque coefficient have been studied for an H-type Darrius vertical axis wind turbine (VAWT). Different types of airfoils were analyzed, and eight of them were selected and divided into two groups. The first group includes the airfoils with camber, including S815, NACA9418, and NACA9415, while the second group including S1048, NACA0018, and NACA0015 have symmetric geometries. The focus of the current study is on two-blade VAWTs because they have higher power coefficient than three or four blades VAWTs. The two-blade VAWTs with selected airfoils were simulated with Computational Fluid Dynamic (CFD) method, and $k-\omega$ SST was used as a turbulence model and then grid independency was checked. The numerical investigation indicates that the cambered airfoils produce a higher static torque coefficient than symmetric ones, up to 79.8%, and are qualified for self-starting purposes. In addition, the symmetric airfoils produce higher power coefficient than cambered ones, up to 68.7%, and are qualified for power extraction purposes.

Keywords Vertical axis wind turbine (VAWT) · Self-starting · Power coefficient · Symmetric airfoils · Cambered airfoils

S. Maleki Dastjerdi (✉) · A. HormoziNejad · K. Gharali
School of Mechanical Engineering, College of Engineering, University of Tehran,
Tehran, Islamic Republic of Iran
e-mail: sajad.maleki@ut.ac.ir

A. HormoziNejad
e-mail: amir.hormozinejad@ut.ac.ir

K. Gharali
e-mail: kgharali@ut.ac.ir; kgharali@uwaterloo.ca

K. Gharali · J. Nathwani
Waterloo Institute for Sustainable Energy (WISE), University of Waterloo,
Waterloo, ON, Canada
e-mail: nathwani@uwaterloo.ca

J. Nathwani
Department of Management Sciences, Department of Civil and Environmental Engineering,
University of Waterloo, Waterloo, ON, Canada

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_35

Nomenclature

U_∞	Freestream velocity (m/s)
D	Turbine diameter (m)
R	Turbine radius (m)
L	Height of the turbine (m)
c	Chord length (m)
λ	Tip speed ratio (–)
$V_{rotation}$	Rotor rotational speed (rad/s)
N	Number of blades (–)
ρ	Density of fluid (kg/m ³)
T	Torque (N m)
C_P	Power coefficient (–)
C_T	Torque coefficient (–)
σ	Solidity (–)

1 Introduction

Lift based vertical axis wind turbines have high power coefficients, but usually, they do not have the self-starting ability at low wind speeds. H-type Darrius vertical axis wind turbines known as the Straight Blade Vertical Axis Wind Turbines (SBVAWTs) are considered as useful types of lift-based VAWTs. Since self-starting is a particular drawback of popular h-darrius VAWTs, many numerical and experimental innovations have been done for improving self-starting ability. SBVAWTs suffer from self-starting issue; that means, the rotor cannot start rotating in itself at low wind velocities. If a wind turbine cannot operate under low wind speeds, it will miss considerable portion of the annual power production. For solving this issue, some turbines are equipped with extra facilities including a motor, controllers and sensors to start rotating at low wind speeds. The extra equipment reduces VAWT simplicity, and causes higher operation and maintenance costs. Thus, designing a low cost VAWT with the self-starting ability without using extra equipment will be remarkable achievement.

Dereng [1] invented a VAWT with a new blade shape. He claimed that it has better self-starting ability than normal VAWTs. Some of the researchers tried to use advantages of both drag-based and lift-based VAWTs, simultaneously. So for achieving this aim, they used blades with flexible shapes, which means they became drag based for self-starting purpose and after starting rotation, the blades are changed to the common type of lift-based airfoils [2]. Batista et al. [3] designed a new airfoil for VAWTs. Their new airfoil had a good self-starting ability. Using a combination of a h-type with a savonius VAWT was another innovation for improving the self-starting ability compared with the SBVAWT [4]; but this innovation resulted in lower power coefficients [4]. Using guide vane for SBVAWT was tried to improve self-starting

ability like drag-based VAWTs [5]. J-shaped airfoils also had positive and negative impacts on self-starting ability and power generation of SBVAWTs, respectively [6, 7]. One of the novel works for improving self-starting ability as well as the power coefficient of SBVAWTs was using auxiliary blades close to the main blades by Li et al. [8] for a two-blade SBVAWT and Scungio et al. [9] for three blade SBVAWT. According to the mentioned studies, the shape of the airfoils has a strong impact on the self-starting ability and the power generation of a SBVAWT. In this study, some airfoils are selected and SBVAWTs with these airfoils are modeled numerically for

- evaluating the effects of cambered airfoils on improving self-starting abilities;
- analyzing the impacts of symmetric airfoils on boosting power coefficients;
- comparing the influences of symmetric airfoils with non-symmetric airfoils on the performance characteristics of VAWTs;
- choosing the best airfoil for improving the self-starting ability;
- selecting the best airfoil for improving the power coefficient; and
- introducing an airfoil with a good self-starting ability and high power coefficient.

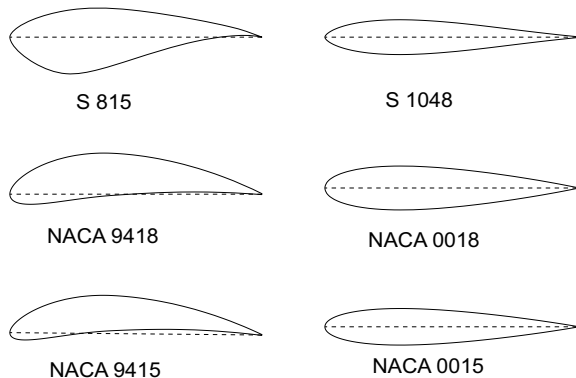
2 Selected Airfoils and Characteristics of VAWT

The selected non-symmetric airfoils are NACA’s airfoils with the high camber, NACA9418, and NACA9415 and also, S815 from Selig’s airfoils. For the symmetric airfoils, NACA0018 and NACA0015 are selected because they have the same thickness ratio as NACA9418 and NACA9415, respectively. S1048 is chosen for comparing with the non-symmetric Selig’s airfoil, S815 (Fig. 1).

The solidity of vertical axis wind turbines is computed by Eq. 1 [10]:

$$\sigma = \frac{N \cdot c}{D} \tag{1}$$

Fig. 1 Selected airfoils



Tip Speed Ratio (TSR) is the ratio of the tip of the blade's velocity to freestream velocity, and it is defined as Eq. 2 [11]:

$$\lambda = \frac{R \cdot V_{rotation}}{U_{\infty}} \quad (2)$$

A two-blade SBVAWT has been simulated numerically. The parameters are set as follows:

- Chord length: 0.15 (m)
- The diameter of the rotor: 1 (m)
- Solidity: 0.3
- H/D: 1
- Tip speed ratios (TSRs): 2, 2.5, 3, 3.5, 4
- The rotational speed of VAWT (constant in different TSRs): 40 (rad/s)
- Free stream velocity for evaluating self-starting ability: 10 (m/s)

3 Methodology

In the current study, both dynamic and static simulations are done with two Intel Xeon E5-2683 V4 (32 cores).

3.1 Setup, Boundary Conditions, and Flow Modeling

For the dynamic simulations, sliding mesh technique under transient condition has been used. Since the velocity of the wind is less than 10 m/s, an incompressible flow and a pressure-based solver have been used. Second-order discretization for spatial discretization and semi-implicit second-order discretization for temporal ones have been applied. For static simulations, the rotating domain is fixed in different angles ranging from 0° to 360° by step size of 30° (Fig. 2).

Reynolds numbers according to the chord length are about 102000 [13]. K- SST has been selected as the turbulence model.

3.2 Mesh

Unstructured meshes are considered for all domains, Fig. 3.

The values of y^+ for all cases are less than four. For evaluating the independency of results from mesh, cells are fined by four steps, and in each level, the number of cells is doubled. The final mesh has 6×10^5 cells (Fig. 4).

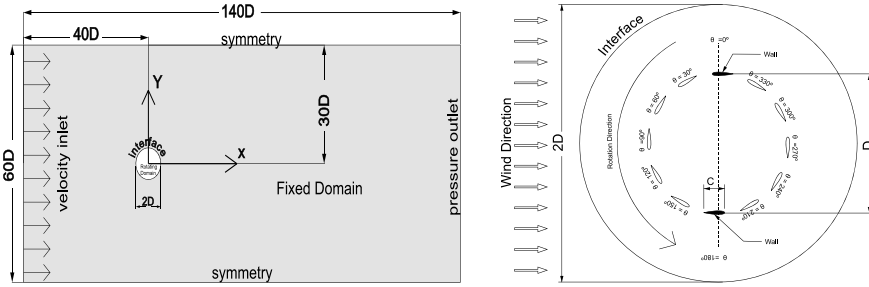


Fig. 2 a Numerical domain and boundary condition (left) [12]; b rotating domain (right)

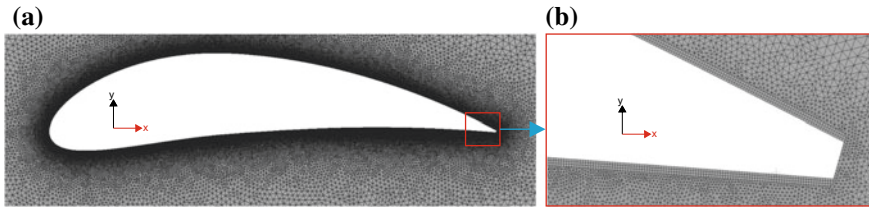


Fig. 3 The 2D mesh; a NACA9418. b Trailing edge of NACA9418

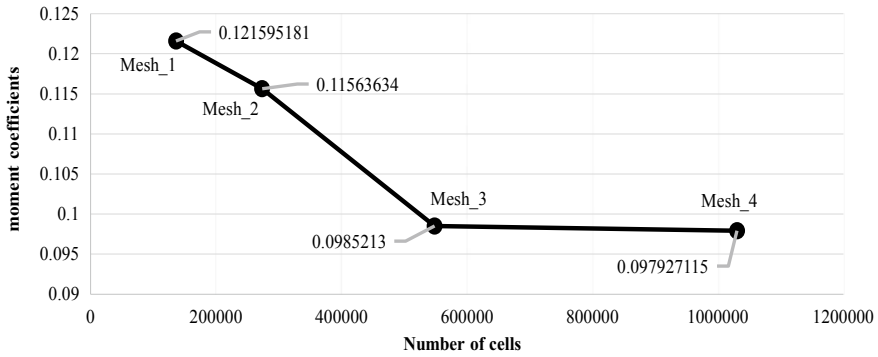


Fig. 4 Grid independency

3.3 Validation

The current results are compared with the numerical and experimental results of Howell et al. [14]. It should be noted that the turbulence model used by Howell et al. was $k-\epsilon$ RNG. The power coefficient extracted from the current study agrees with the experimental data, Fig. 5.

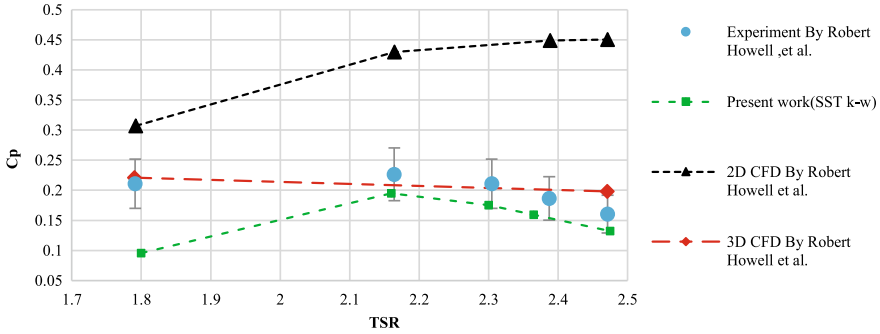


Fig. 5 Validation and verification of simulation

4 Evaluation of Power Coefficients (Cp)

The average static torque coefficient of the wind turbine is obtained by

$$C_T = \frac{T}{\frac{1}{4} \cdot \rho (U_\infty)^2 \cdot D^2 \cdot L} \tag{3}$$

And the power coefficient, C_p is calculated from $\frac{C_p}{C_T} = \lambda$. The combination of the pressure and the viscous forces of the airfoils results in torque calculation. The pressure force is determined by integrating the pressure values around the airfoil. The viscous forces are measured through the boundary layer of airfoils. Finally, the tangential component of the obtained pressure and viscous forces is multiplied by the radius of the rotor to calculate the torque value. Then, the torque coefficient (C_T) will be computed by Eq. 3.

From Figs. 6 and 7, it can be concluded that although NACA9418 is a non-symmetric airfoil, it behaves as symmetric airfoils in terms of power coefficient variation. The airfoil S815 is not shown for TSRs bigger than two since the average power coefficient of this airfoil is negative in these TSRs.

The maximum power coefficient of 0.37 is for NACA0015 at TSR=2.5, Fig. 8a. For higher wind speeds, NACA0018 and NACA0015 are more efficient than other airfoils. For lower wind speeds, S1048 has the best power coefficient. The power coefficient of NACA9418 for TSR of 3.5 is more than the power coefficient of NACA0018. Therefore, for lower wind speeds, NACA9418 is more efficient.

In real cases, the wind speed is not constant. Therefore, the average power coefficients in all TSRs have been plotted in Fig. 8b for the wind speed range from 5 to 10 m/s. Although the average power coefficients of symmetric airfoils are more than non-symmetric ones, NACA9418 has higher C_p than S1048.

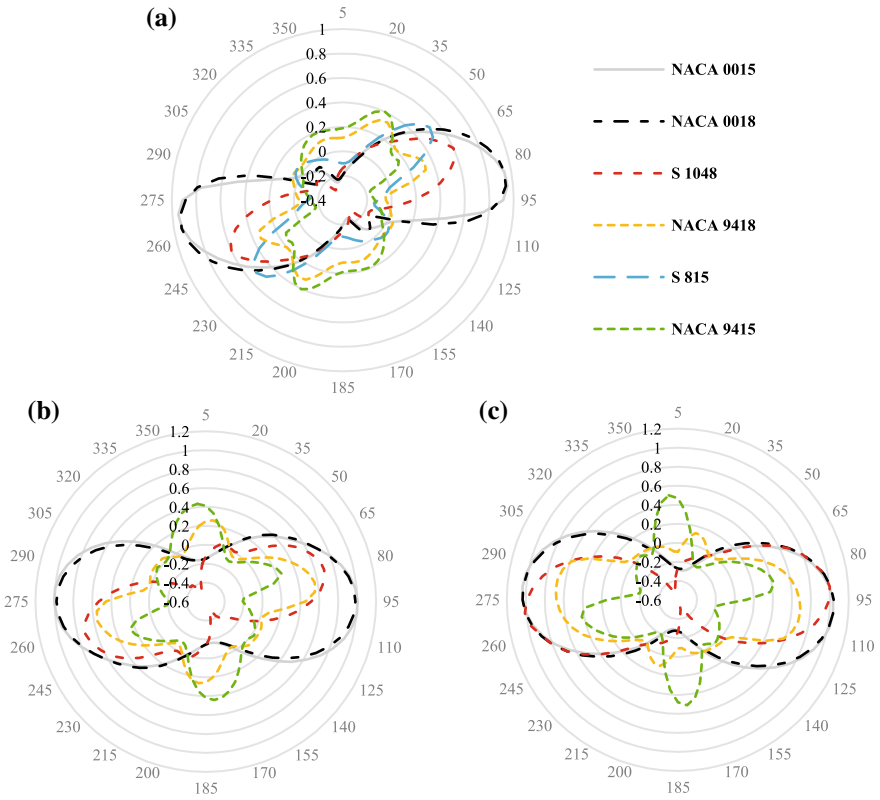


Fig. 6 Power coefficients for TSRs lower than 3: a TSR = 2; b TSR = 2.5; c TSR = 3

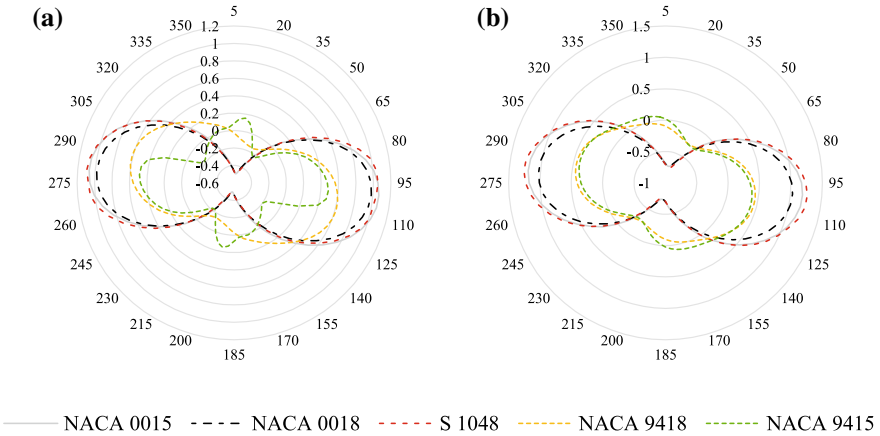


Fig. 7 Power coefficient for TSRs higher than 3: a TSR = 3.5; b TSR = 4

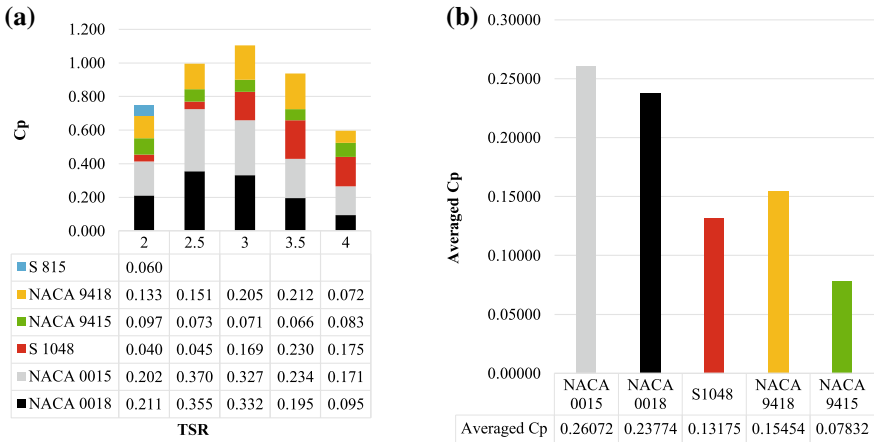


Fig. 8 a C_p versus TSR; b average power coefficients of different airfoils for all evaluated TSRs

5 Analysis of Self-starting

The static torque coefficients for all six airfoils except NACA9415 and NACA9418 are negative at 90° and 270° , Table 1. The arrangement of the blades at 90° and 270° in front of the wind makes the blades like a bluff body.

S815 behaves like a symmetric airfoil. S815 and NACA9418 have the highest self-starting ability from 0° to 60° and from 90° to 150° , respectively.

The average static torque coefficients at wind speed of 10m/s are plotted in Fig. 9a. The non-symmetric airfoils have higher self-starting ability than the symmetric ones.

Table 1 Static torque coefficients

Angle ($^\circ$)	NACA 0018	NACA 9418	NACA 0015	NACA 9415	S 1048	S 815
0	0.00610	0.01448	0.00323	0.00955	0.00495	0.01506
30	0.04275	0.00132	0.03523	0.00021	0.02906	0.05042
60	0.02262	0.00500	0.02530	0.00381	0.02136	0.02934
90	-0.01368	0.00677	-0.01711	0.00296	-0.01467	-0.02570
120	0.00496	0.03615	0.00406	0.03382	-0.00205	-0.00056
150	0.00410	0.05645	0.00709	0.04956	0.00491	0.02534
180	0.00610	0.01448	0.00323	0.00955	0.00495	0.01506
210	0.04275	0.00132	0.03523	0.00021	0.02906	0.05042
240	0.02262	0.00500	0.02530	0.00381	0.02136	0.02934
270	-0.01368	0.00677	-0.01711	0.00296	-0.01467	-0.02570
300	0.00496	0.03615	0.00406	0.03382	-0.00205	-0.00056
330	0.00410	0.05645	0.00709	0.04956	0.00491	0.02534

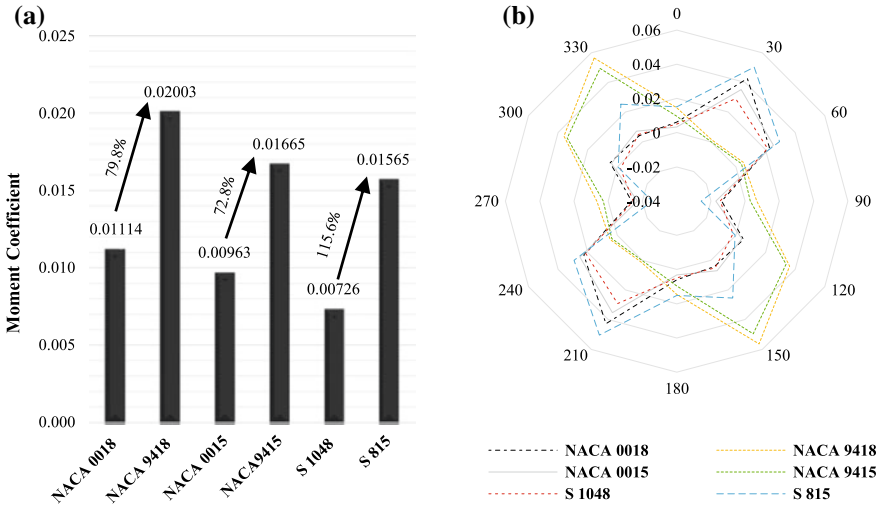


Fig. 9 **a** Average static torque coefficients; **b** details of static torque coefficients in one rotation

Among the non-symmetric and symmetric airfoils, NACA9418 and NACA0018 have the highest average static torque coefficients; but S815 and S1048 have the lowest ones. The static torque coefficient of NACA9418 is about 79.8% and 108% more than that of NACA0018 and NACA0015, respectively.

6 Conclusion

2D numerical simulations have been done for comparing selected non-symmetric airfoils versus symmetric airfoils based on the power coefficient and self-starting ability of a vertical axis wind turbine. The results can help airfoil selection for VAWT design. In general, non-symmetric airfoils have better self-starting ability than symmetric airfoils. However, symmetric airfoils have a higher power coefficient. So if self-starting is not a crucial issue, a symmetric airfoil is a better choice for a VAWT.

According to the current study based on the selected airfoils, NACA0015 and NACA0018 have 68.7 and 53.8% more power coefficients than NACA9418 which has the highest power coefficient among the rest of the non-symmetric airfoils. NACA0015 has the highest power coefficient, so Ignoring the self-starting ability, NACA0015 can be considered as a suitable choice for a VAWT. NACA9418, NACA9415, and S815 have 79.8%, 49.4%, and 40.5% higher self-starting ability, respectively than NACA0018, which has the highest self-starting ability among the rest of the symmetric airfoils. NACA9418 increases the self-starting ability of the turbine while decreases the power coefficient less than the other non-symmetric airfoils. If both self-starting and power coefficient are important, NACA9418 might

be assumed as the best choice. The non-symmetric S815 generates negative power coefficient for some conditions while it improves slightly the self-starting ability. Although S815 is a non-symmetric airfoil, it behaves like symmetric airfoils due to its high thickness and low camber.

Thicker airfoils have better self-starting ability according to this study. Then, NACA0018 has a better self-starting ability than NACA0015, and NACA9418 is better than NACA9415. S1048 airfoil has the worst self-starting ability among the rest of the airfoils. It generates the lowest power among the symmetric airfoils, even lower than non-symmetric NACA9418. However, in low wind speeds with TSR of four, it produces the highest power, so it can be concluded that performance of S1048 extremely depends on the wind speed.

References

1. Dereng, V.G.: Fixed geometry self starting transverse axis wind turbine. US Patent 4,264,279 (1981)
2. Manohar, K., Rampartap, A., Ramkissoon, R.: Self-starting hybrid “h” type wind turbine. In: ASME 2007 Energy Sustainability Conference, pp. 1139–1146. American Society of Mechanical Engineers Digital Collection (2009)
3. Batista, N., Melicio, R., Matias, J., Catalão, J.: New blade profile for darrieus wind turbines capable to self-start (2011)
4. Mohamed, M.: Impacts of solidity and hybrid system in small wind turbines performance. *Energy* **57**, 495–504 (2013)
5. Siregar, I.H., Ansori, A.: Performance of combined vertical axis wind turbine blade between airfoil NACA 0018 with curve blade with and without guide vane. Jurusan Teknik Mesin Fakultas Teknik Universitas Surabaya, Surabaya (2016)
6. Chen, J., Yang, H., Yang, M., Xu, H.: The effect of the opening ratio and location on the performance of a novel vertical axis Darrieus turbine. *Energy* **89**, 819–834 (2015)
7. Zamani, M., Nazari, S., Moshizi, S.A., Maghrebi, M.J.: Three dimensional simulation of j-shaped Darrieus vertical axis wind turbine. *Energy* **116**, 1243–1255 (2016)
8. Li, Y., Zhao, C., Qu, C., Zhao, S., Feng, F., Tagawa, K.: Effect of auxiliary blade on aerodynamic characteristics of vertical axis wind turbine by numerical simulation. *Int. J. Rotating Mach.* **2019** (2019)
9. Scungio, M., Arpino, F., Focanti, V., Profili, M., Rotondi, M.: Wind tunnel testing of scaled models of a newly developed Darrieus-style vertical axis wind turbine with auxiliary straight blades. *Energy Convers. Manag.* **130**, 60–70 (2016)
10. Fiedler, A.J., Tullis, S.: Blade offset and pitch effects on a high solidity vertical axis wind turbine. *Wind Eng.* **33**(3), 237–246 (2009)
11. Ferreira, C.S., Van Kuik, G., Van Bussel, G., Scarano, F.: Visualization by PIV of dynamic stall on a vertical axis wind turbine. *Exp. Fluids* **46**(1), 97–108 (2009)
12. Bianchini, A., Balduzzi, F., Ferrara, G., Ferrari, L.: Influence of the blade-spoke connection point on the aerodynamic performance of Darrieus wind turbines. In: ASME Turbo Expo 2016: Turbomachinery Technical Conference and Exposition. American Society of Mechanical Engineers Digital Collection (2016)
13. Fox, R.W., McDonald, A.T., Pritchard, P.: Introduction to Fluid Mechanics, 5th edn. (2010)
14. Howell, R., Qin, N., Edwards, J., Durrani, N.: Wind tunnel and numerical study of a small vertical axis wind turbine. *Renew. Energy* **35**(2), 412–422 (2010)

Stability and Stabilization of T–S Fuzzy Systems with Aperiodic Sampling



Jinnan Luo, Xinzhi Liu, Wenhong Tian, Shouming Zhong, and Kaibo Shi

Abstract This study aims at investigating the problem of stability and stabilization of chaotic systems on the basis of Takagi–Sugeno (T–S) fuzzy model under aperiodic sampling. A modified Lyapunov–Krasovskii functional (LKF), which fully captures the information of the sampling pattern, is constructed to the chaotic systems. Together with free-weighted matrices technique, much less conservative stabilization results are derived in term of linear matrix inequality (LMI).

Keywords Chaotic systems · Stability and stabilization · Aperiodic sampling · T-S fuzzy model

J. Luo (✉)

School of Electronic Engineering, Southwest University for Nationalities, Chengdu 610041, People's Republic of China

e-mail: jinnanluo@outlook.com

W. Tian

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

e-mail: tian_wenhong@uestc.edu.cn

X. Liu

Department of Applied Mathematics, University of Waterloo, Waterloo N2L 3G1, Canada

e-mail: xinzhi.liu@uwaterloo.ca

S. Zhong

School of Mathematics Sciences, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

e-mail: zhongsm@uestc.edu.cn

K. Shi

School of Information Science and Engineering, Chengdu University, Chengdu 610106, People's Republic of China

e-mail: skbs111@163.com

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_36

393

1 Introduction

T–S fuzzy systems represented by a batteries of IF-THEN rules with fuzzy sets, which offers an efficient tool to depict complex nonlinear systems, have been extensively studied [1–3]. Recently, plentiful results have been investigated the T–S fuzzy systems [4–8].

With the development of microelectronics, increasing attention has been drawn to the sampled-data control, which greatly improve the efficiency of the control. It does not require to transmit the information moment by moment. Hence, numerous results have been studied the sampled-data control [9–12]. The authors in [13] have researched stabilization of the fuzzy system with sampled-data control by the switched operation approach. The H_∞ stabilization results have been presented for T–S fuzzy systems by memory sampled-data control in [14]. In [15], the stabilization of chaotic systems based on a T–S fuzzy model has been studied by the sampled-data control.

In this study, stability and stabilization of chaotic systems on the basis of T–S fuzzy model via aperiodic sampling is investigated. The core contributions of this study are presented as aspects: (1) The stabilization for T–S fuzzy systems with aperiodic sampling is researched. (2) A modified LKF is constructed and less conservative results are obtained. (3) The sampled-data controller is devised for the system.

Notations: $R^{n \times m}$ is the set of all $n \times m$ real matrices, R^m denotes m -dimensional Euclidean space. $diag\{.\dots\}$ denotes block diagonal matrix, $\mathcal{A} < (>) \mathcal{O}$ denoting \mathcal{A} is the negative (positive) definite matrix, $O_{n \times n}$ stands for the $n \times n$ zero matrix, I_n denotes the $n \times n$ identity matrix, $*$ is the symmetric block in symmetric matrix and $He\{\mathcal{C}\} = \mathcal{C} + \mathcal{C}^T$.

2 Preliminaries

Consider the fuzzy model as below:

Plant rule j : IF $\mathbb{h}_1(t)$ is \mathbb{x}_{j1} , $\mathbb{h}_2(t)$ is \mathbb{x}_{j2} , \dots and $\mathbb{h}_r(t)$ is \mathbb{x}_{jr} , THEN

$$\dot{\mathbb{z}}(t) = \mathcal{B}_j \mathbb{z}(t) + \mathcal{D}_j u(t), \tag{1}$$

where $\mathbb{z}(t) \in R^n$ denotes the state, $u(t) \in R^\zeta$ is a control input. $\mathbb{h}_1(t)$, $\mathbb{h}_2(t)$, \dots , $\mathbb{h}_r(t)$ denote premise variables, \mathbb{x}_{ji} represents fuzzy sets, $j = 1, 2, \dots, \alpha$, $i = 1, 2, \dots, r$, \mathcal{B}_j and \mathcal{D}_j are given matrices with proper dimensions, α stands for the number of IF-THEN rules.

$$\text{Set } \mathbb{h}(t) = [\mathbb{h}_1(t), \mathbb{h}_2(t), \dots, \mathbb{h}_r(t)], \quad \mathbb{q}_j(\mathbb{h}(t)) = \frac{\prod_{i=1}^r \mathbb{x}_{ji}(\mathbb{h}_i(t))}{\sum_{j=1}^\alpha \prod_{i=1}^r \mathbb{x}_{ji}(\mathbb{h}_i(t))} \text{ with } \mathbb{x}_{ji}(\mathbb{h}_i(t))$$

is the grade of membership of $\mathbb{h}_i(t)$ in \mathbb{x}_{ji} and suppose that $\prod_{i=1}^r \mathbb{x}_{ji}(\mathbb{h}_i(t)) \geq 0$,

$$\sum_{j=1}^{\alpha} \prod_{i=1}^r \mathfrak{x}_{ji}(\mathfrak{h}_i(t)) > 0, \forall t > 0,$$

$$\mathfrak{q}_j(\mathfrak{h}(t)) \geq 0, \quad \sum_{j=1}^{\alpha} \mathfrak{q}_j(\mathfrak{h}(t)) = 1. \tag{2}$$

By fuzzy blending, we have the following fuzzy system (3)

$$\dot{\mathfrak{z}}(t) = \sum_{j=1}^{\alpha} \mathfrak{q}_j(\mathfrak{h}(t))[\mathcal{B}_j \mathfrak{z}(t) + \mathcal{D}_j u(t)]. \tag{3}$$

The control input is given in the following

Rule i : IF $\mathfrak{h}_1(t)$ is \mathfrak{x}_{i1} , $\mathfrak{h}_2(t)$ is \mathfrak{x}_{i2} , ... and $\mathfrak{h}_r(t)$ is \mathfrak{x}_{ir} , THEN

$$u(t) = \mathcal{K}_i \mathfrak{z}(t - \mathfrak{s}(t)), \tag{4}$$

where $\mathfrak{s}(t) = t - t_k$. The control signal is produced by zero-order hold with a battery of hold times

$$0 = t_0 < t_1 < \dots < t_k < \dots < \lim_{k \rightarrow \infty} t_k = +\infty, \tag{5}$$

where $t_k \leq t < t_{k+1}$, \mathcal{K}_i is the control gain and $\mathfrak{z}(t_k)$ is the measurement of $\mathfrak{z}(t)$ at the sampling instant t_k . Assume that $0 < \mathfrak{s}(t) \leq t_{k+1} - t_k = \mathfrak{s}_k \leq \mathfrak{s}, \forall k \geq 0$. Then, we obtain the following controller (6)

$$u(t) = \sum_{i=1}^{\alpha} \mathfrak{q}_i(\mathfrak{h}(t - \mathfrak{s}(t))) \mathcal{K}_i \mathfrak{z}(t - \mathfrak{s}(t)), \quad t_k \leq t < t_{k+1}. \tag{6}$$

Replacing (6) with (3), we get

$$\dot{\mathfrak{z}}(t) = \sum_{j=1}^{\alpha} \sum_{i=1}^{\alpha} \mathfrak{q}_j(\mathfrak{h}(t)) \mathfrak{q}_i(\mathfrak{h}(t - \mathfrak{s}(t))) [\mathcal{B}_j \mathfrak{z}(t) + \mathcal{D}_j \mathcal{K}_i \mathfrak{z}(t - \mathfrak{s}(t))]. \tag{7}$$

3 Main Results

In this part, we will present the stability and stabilization results for the system (7). For convenience, let $\mathfrak{v}(t) = [\mathfrak{z}^T(t) \quad \mathfrak{z}^T(t - \mathfrak{s}(t)) \quad \dot{\mathfrak{z}}^T(t) \quad \mathfrak{z}^T(t - \mathfrak{s}) \int_{t-\mathfrak{s}}^t \mathfrak{z}^T(\alpha) d\alpha]^T$, $\bar{\mathfrak{v}}(t) = [\mathfrak{z}^T(t) \quad \mathfrak{z}^T(t - \mathfrak{s}(t)) \quad \dot{\mathfrak{z}}^T(t) \quad \mathfrak{z}^T(t - \mathfrak{s}) \int_{t-\mathfrak{s}}^t \mathfrak{z}^T(\alpha) d\alpha]^T$, $\mathbb{E}_i = [O_{n \times (i-1)n} \quad \mathbb{I}_n \quad O_{n \times (5-i)n}]$, $i = 1, 2, \dots, 5$.

Theorem 1 For a given scalar $s > 0$ and control gains \mathcal{K}_i , the system (7) is asymptotically stable if there exist matrices $\mathcal{P} > O$, $\mathcal{Y} > O$, $\mathcal{W}_1 > O$, $\mathcal{W}_2 > O$, $\mathcal{X} > O$ and any matrices $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ such that the following LMI hold for $i, j = 1, 2, \dots, \alpha$

$$\tilde{\Upsilon}_{ij} < O, \tag{8}$$

where

$$\begin{aligned} \tilde{\Upsilon}_{ij} = & \text{He}\{[\mathbb{E}_1 \ \mathbb{E}_5]^T \mathcal{P} [\mathbb{E}_3 \ \mathbb{E}_1 - \mathbb{E}_4]\} + \mathbb{E}_1^T (\mathcal{Y} + s\mathcal{W}_2)\mathbb{E}_1 - \mathbb{E}_4^T \mathcal{Y}\mathbb{E}_4 \\ & + \mathbb{E}_3^T (s\mathcal{W}_1 + \frac{s^2}{2}\mathcal{X})\mathbb{E}_3 - \frac{1}{s}(\mathbb{E}_1 - \mathbb{E}_2)^T \mathcal{W}_1(\mathbb{E}_1 - \mathbb{E}_2) \\ & - \frac{1}{s}\mathbb{E}_5^T \mathcal{W}_2\mathbb{E}_5 - \frac{2}{s^2}(\mathbb{E}_1 - \mathbb{E}_5)^T \mathcal{X}(\mathbb{E}_1 - \mathbb{E}_5) \\ & + \text{He}\{[\mathbb{E}_1^T \mathcal{F}_1 + \mathbb{E}_3^T \mathcal{F}_2 + \mathbb{E}_2^T \mathcal{F}_3] \\ & \times [-\mathbb{E}_3 + \mathcal{B}_j\mathbb{E}_1 + \mathcal{D}_j\mathcal{K}_i\mathbb{E}_2]\}. \end{aligned}$$

Proof Consider the LKF as

$$V(t) = \sum_{c=1}^4 V_c(t), \quad t \in [t_k, t_{k+1}), \tag{9}$$

with

$$V_1(t) = \left[\int_{t-s}^t z(t) \right]^T \mathcal{P} \left[\int_{t-s}^t z(\alpha) d\alpha \right],$$

$$V_2(t) = \int_{t-s}^t z^T(\alpha) \mathcal{Y} z(\alpha) d\alpha,$$

$$\begin{aligned} V_3(t) = & \int_{-s}^0 \int_{t+x}^t \dot{z}^T(\alpha) \mathcal{W}_1 \dot{z}(\alpha) d\alpha d\mathbb{x} \\ & + \int_{-s}^0 \int_{t+x}^t z^T(\alpha) \mathcal{W}_2 z(\alpha) d\alpha d\mathbb{x}, \end{aligned}$$

$$V_4(t) = \int_{-s}^0 \int_{\mathbb{x}} \int_{t+\alpha}^t \dot{z}^T(\beta) \mathcal{X} \dot{z}(\beta) d\beta d\alpha d\mathbb{x}.$$

Calculating the derivative of $V_c(t)$, $c = 1, 2, \dots, 4$ along with the system (7), we get

$$\dot{V}_1(t) = \text{He}\left\{ \left[\int_{t-s}^t z(t) \right]^T \mathcal{P} \left[\begin{matrix} \dot{z}(t) \\ z(t) - z(t-s) \end{matrix} \right] \right\}, \tag{10}$$

$$\dot{V}_2(t) = z^T(t) \mathcal{Y} z(t) - z^T(t-s) \mathcal{Y} z(t-s), \tag{11}$$

$$\begin{aligned} \dot{V}_3(t) &= \mathfrak{s} \dot{z}^T(t) \mathcal{W}_1 \dot{z}(t) - \int_{t-\mathfrak{s}}^t \dot{z}^T(\alpha) \mathcal{W}_1 \dot{z}(\alpha) d\alpha \\ &\quad + \mathfrak{s} z^T(t) \mathcal{W}_2 z(t) - \int_{t-\mathfrak{s}}^t z^T(\alpha) \mathcal{W}_2 z(\alpha) d\alpha, \end{aligned} \quad (12)$$

$$\dot{V}_4(t) = \frac{\mathfrak{s}^2}{2} \dot{z}^T(t) \mathcal{X} \dot{z}(t) - \int_{-\mathfrak{s}}^0 \int_{t+\mathfrak{x}}^t \dot{z}^T(\alpha) \mathcal{X} \dot{z}(\alpha) d\alpha d\mathfrak{x}. \quad (13)$$

By dealing with the integral terms in (12) and (13), we have

$$- \int_{t-\mathfrak{s}}^t \dot{z}^T(\alpha) \mathcal{W}_1 \dot{z}(\alpha) d\alpha \leq -\frac{1}{\mathfrak{s}} (\mathbb{E}_1 - \mathbb{E}_2)^T \mathcal{W}_1 (\mathbb{E}_1 - \mathbb{E}_2), \quad (14)$$

$$- \int_{t-\mathfrak{s}}^t z^T(\alpha) \mathcal{W}_2 z(\alpha) d\alpha \leq -\frac{1}{\mathfrak{s}} \mathbb{E}_5^T \mathcal{W}_2 \mathbb{E}_5, \quad (15)$$

$$- \int_{-\mathfrak{s}}^0 \int_{t+\mathfrak{x}}^t \dot{z}^T(\alpha) \mathcal{X} \dot{z}(\alpha) d\alpha d\mathfrak{x} \leq -\frac{2}{\mathfrak{s}^2} (\mathbb{E}_1 - \mathbb{E}_5)^T \mathcal{X} (\mathbb{E}_1 - \mathbb{E}_5). \quad (16)$$

From the system (7), we derive

$$\begin{aligned} 0 &= \sum_{j=1}^{\alpha} \sum_{i=1}^{\alpha} \mathfrak{q}_j(\mathfrak{h}(t)) \mathfrak{q}_i(\mathfrak{h}(t - \mathfrak{s}(t))) \text{He}\{[z^T(t) \mathcal{F}_1 \\ &\quad + \dot{z}^T(t) \mathcal{F}_2 + z^T(t - \mathfrak{s}(t)) \mathcal{F}_3] \\ &\quad \times [-\dot{z}(t) + \mathcal{B}_j z(t) + \mathcal{D}_j \mathcal{K}_i z(t - \mathfrak{s}(t))]\}. \end{aligned} \quad (17)$$

By combining (10)–(16) with (17), we obtain

$$\dot{V}(t) \leq \sum_{j=1}^{\alpha} \sum_{i=1}^{\alpha} \mathfrak{q}_j(\mathfrak{h}(t)) \mathfrak{q}_i(\mathfrak{h}(t - \mathfrak{s}(t))) \mathfrak{v}^T(t) \Upsilon \mathfrak{v}(t), \quad (18)$$

where Υ is defined in Theorem 1.

According to the condition (8), when $\mathfrak{v}(t) \neq O$, it follows that

$$\dot{V}(t) < 0, \quad (19)$$

Therefore, the system (7) is asymptotically stable.

Theorem 1 is based on the given control gain \mathcal{K}_i . The following Theorem 2 will solve the control gain if it is not given in advance.

Theorem 2 For given scalars $\mathfrak{s} > 0$, ψ_1, ψ_2 , the system (7) is asymptotically stable if there exist matrices $\hat{\mathcal{P}} > O$, $\hat{\mathcal{Y}} > O$, $\hat{\mathcal{W}}_1 > O$, $\hat{\mathcal{W}}_2 > O$, $\hat{\mathcal{X}} > O$ and any matrices $\mathcal{T}, \mathcal{Q}_i$ such that the following LMI hold for $i, j = 1, 2, \dots, \alpha$

$$\hat{Y}_{ij} < O, \tag{20}$$

where

$$\begin{aligned} \hat{Y}_{ij} = & \text{He}\{[\mathbb{E}_1 \ \mathbb{E}_5]^T \hat{\mathcal{P}} [\mathbb{E}_3 \ \mathbb{E}_1 - \mathbb{E}_4]\} + \mathbb{E}_1^T (\hat{\mathcal{Y}} + s\hat{\mathcal{W}}_2)\mathbb{E}_1 - \mathbb{E}_4^T \hat{\mathcal{Y}}\mathbb{E}_4 \\ & + \mathbb{E}_3^T (s\hat{\mathcal{W}}_1 + \frac{s^2}{2}\hat{\mathcal{X}})\mathbb{E}_3 - \frac{1}{s}(\mathbb{E}_1 - \mathbb{E}_2)^T \hat{\mathcal{W}}_1(\mathbb{E}_1 - \mathbb{E}_2) \\ & - \frac{1}{s}\mathbb{E}_5^T \hat{\mathcal{W}}_2\mathbb{E}_5 - \frac{2}{s^2}(\mathbb{E}_1 - \mathbb{E}_5)^T \hat{\mathcal{X}}(\mathbb{E}_1 - \mathbb{E}_5) \\ & + \text{He}\{[\mathbb{E}_1^T \mathbb{I}_n + \mathbb{E}_3^T \psi_1 \mathbb{I}_n + \mathbb{E}_2^T \psi_2 \mathbb{I}_n] \\ & \times [-\mathcal{T}^T \mathbb{E}_3 + \mathcal{B}_j \mathcal{T}^T \mathbb{E}_1 + \mathcal{D}_j \mathcal{L}_i \mathbb{E}_2]\}, \end{aligned}$$

The control gain is designed as

$$\mathcal{K}_i = \mathcal{L}_i(\mathcal{T}^T)^{-1}. \tag{21}$$

Proof Denote $\mathcal{F}_1 = \mathcal{T}^{-1}$, $\mathcal{F}_2 = \psi_1 \mathcal{T}^{-1}$, $\mathcal{F}_3 = \psi_2 \mathcal{T}^{-1}$, $\hat{\mathcal{P}} = \mathcal{T} \mathcal{P} \mathcal{T}^T$, $\hat{\mathcal{Y}} = \mathcal{T} \mathcal{Y} \mathcal{T}^T$, $\hat{\mathcal{W}}_1 = \mathcal{T} \mathcal{W}_1 \mathcal{T}^T$, $\hat{\mathcal{W}}_2 = \mathcal{T} \mathcal{W}_2 \mathcal{T}^T$, $\hat{\mathcal{X}} = \mathcal{T} \mathcal{X} \mathcal{T}^T$, $\Gamma = \text{diag}\{\mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}\}$, $\mathcal{L}_i = \mathcal{K}_i \mathcal{T}^T$. Pre and postmultiplying (8) with Γ , Γ^T , we have (20).

4 Conclusion

We have investigated the problems of stability and stabilization of T-S fuzzy model under aperiodic sampling. An improved LKF has been considered and the aperiodic sampling methods are utilized. Based on the free-weighted matrices technique, much less conservative stability and stabilization results are derived in term of LMI. Future works will aim at the stabilization of Lurie systems, memristor-based neural networks, multi-agent systems by state quantized control and impulsive control.

References

1. Cheng, J., Park, J.H., Zhang, L., Zhu, Y.: An asynchronous operation approach to event-triggered control for fuzzy Markovian jump systems with general switching policies. *IEEE Trans. Fuzzy Syst.* **26**(1), 6–18 (2018)
2. Wang, X., Park, J.H., She, K., Zhong, S., Shi, L.: Stabilization of chaotic systems with T-S fuzzy model and nonuniform sampling: A switched fuzzy control approach. *IEEE Trans. Fuzzy Syst.* **27**(6), 1263–1271 (2019)
3. Zhang, R., Liu, X., Zeng, D., Zhong, S., Shi, K.: A novel approach to stability and stabilization of fuzzy sampled-data Markovian chaotic systems. *Fuzzy Sets Syst.* **344**, 108–128 (2018)
4. Feng, Z., Zheng, W.: Improved stability condition for Takagi-Sugeno fuzzy systems with time-varying delay. *IEEE Trans. Cybern.* **47**(3), 661–670 (2017)

5. Liu, Y., Guo, B.-Z., Park, J.H., Lee, S.: Event-based reliable dissipative filtering for T-S fuzzy systems with asynchronous constraints. *IEEE Trans. Fuzzy Syst.* **26**(4), 2089–2098 (2018)
6. Luo, J., Li, M., Liu, X., Tian, W., Zhong, S., Shi, K.: Stabilization analysis for fuzzy systems with a switched sampled-data control. *J. Frankl. Inst.* **357**(1), 39–58 (2020)
7. Li, Y.-X., Yang, G.-H.: Observer-based adaptive fuzzy quantized control of uncertain nonlinear systems with unknown control directions. *Fuzzy Sets Syst.* **371**, 61–77 (2019)
8. Cheng, J., Zhang, D., Qi, W., Cao, J., Shi, K.: Finite-time stabilization of T-S fuzzy semi-Markov switching systems: a coupling memory sampled-data control approach. *J. Frankl. Inst.* 1–16 (2019). <https://doi.org/10.1016/j.jfranklin.2019.06.021>
9. Liu, Y., Park, J.H., Guo, B., Shu, Y.: Further results on stabilization of chaotic systems based on fuzzy memory sampled-data control. *IEEE Trans. Fuzzy Syst.* **26**(2), 1040–1045 (2018)
10. Wang, Y., Xia, Y., Zhou, P.: Fuzzy-model-based sampled-data control of chaotic systems: a fuzzy time-dependent Lyapunov-Krasovskii functional approach. *IEEE Trans. Fuzzy Syst.* **25**(6), 1672–1684 (2017)
11. Li, S., Ahn, C.K., Xiang, Z.: Sampled-data adaptive output feedback fuzzy stabilization for switched nonlinear systems with asynchronous switching. *IEEE Trans. Fuzzy Syst.* **27**(1), 200–205 (2018)
12. Lee, T.H., Park, J.H.: Stability Analysis of sampled-data systems via free-matrix-based time-dependent discontinuous Lyapunov approach. *IEEE Trans. Autom. Control* **62**(7), 3653–3657 (2017)
13. Wang, X., Park, J.H., Zhong, S., Yang, H.: A switched operation approach to sampled-data control stabilization of fuzzy memristive neural networks with time-varying delay. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(1), 891–900 (2020)
14. Ge, C., Shi, Y., Park, J.H., Hua, C.: Robust H_∞ stabilization for T-S fuzzy systems with time-varying delays and memory sampled-data control. *Appl. Math. Comput.* **346**, 500–512 (2019)
15. Zeng, H.-B., Teo, K.L., He, Y., Wang, W.: Sampled-data stabilization of chaotic systems based on a T-S fuzzy model. *Inf. Sci.* **483**, 262–272 (2019)

A New Method of Modelling Tuneable Lasers with Functional Composition



B. Metherall and C. Sean Bohun

Abstract A new nonlinear model is proposed for tuneable lasers. Using the generalized nonlinear Schrödinger equation as a starting point, expressions for the transformations undergone by the pulse are derived for each of the five components (gain, loss, dispersion, modulation, and nonlinearity) within the laser cavity. These transformations are then composed to give the overall effect of one trip around the cavity. This is in contrast to solving the generalized nonlinear Schrödinger equation which treats the processes as continuous.

Keywords Tuneable ring lasers · Modulation instability · Mathematical modelling · Generalized nonlinear Schrödinger equation

1 Introduction

A tuneable laser has the ability to vary the frequency of its output by up to about 100 nm [5, 8, 38]. Tuneable lasers simultaneously lase at all frequencies within this bandwidth. This tuneability is quite useful and has applications in spectroscopy and high resolution imaging such as coherent anti-Stokes Raman spectroscopy and optical coherence tomography [5, 7, 38], as well as communications and diagnostics of ultra fast processes [31]. A typical tuneable laser cavity can be seen in Fig. 1. In contrast to a standard laser, a tuneable laser contains two additional components, namely, a chirped fibre Bragg grating (CFBG), and a modulator.

A CFBG is a length of optical fibre where the refractive index oscillates along its length [10], and therefore, can act as a reflective filter [1, 3, 10, 32]. Due to the

B. Metherall (✉)

University of Oxford, Radcliffe Observatory Quarter, Andrew Wiles Building, Woodstock Rd, Oxford OX2 6GG, UK

e-mail: brady.metherall@maths.ox.ac.uk

C. S. Bohun

Ontario Tech University, 2000 Simcoe St N, Oshawa, ON L1G 0C5, Canada

e-mail: sean.bohun@uoit.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_37

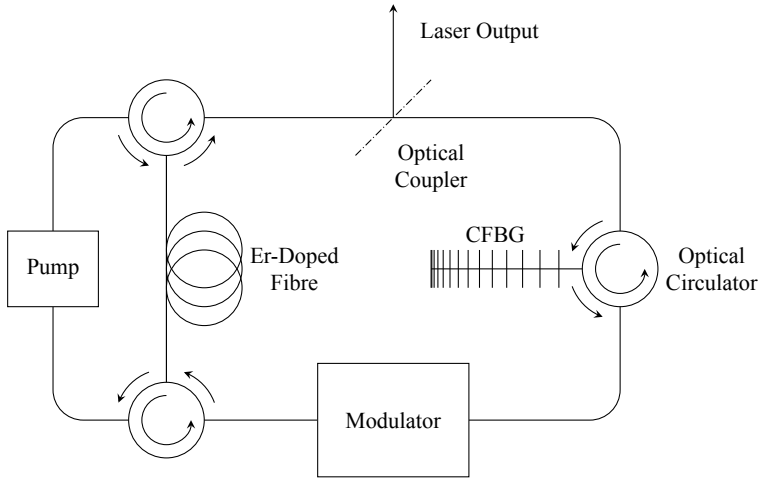


Fig. 1 Typical cavity of a fibre based tuneable laser. The laser pulses travel clockwise around each loop. The pulses iteratively pass through each component successively

oscillatory nature, light with the corresponding wavelength will be reflected when the Bragg condition is satisfied [1, 3, 4, 10, 31, 32]. The spacial variation of the refractive index effectively creates a spacial dependence on the Bragg condition, causing most wavelengths to be reflected by a CFBG, but with each wavelength satisfying the Bragg condition at a different spacial location.¹ A consequence of this is that a time delay is created between wavelengths—this causes the pulse to disperse and broaden.

The modulator serves the purpose of reshaping the pulse. Without it, the pulse will repeatedly widen due to the CFBG—the modulator ensures the pulse is band limited by altering the envelope.

2 Previous Modelling Efforts

The standard equation for studying nonlinear optics is the nonlinear Schrödinger equation (NLSE),²

$$\frac{\partial A}{\partial z} = -i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2} + i\gamma |A|^2 A. \quad (1)$$

¹ Note that a monotonic chirping ensures that the spacial dependence of the Bragg condition is continuous with respect to the frequency.

² The NLSE can be derived from the nonlinear wave equation for electric fields; this derivation is presented in detail in [2, 10].

Here $A = A(T, z) : \mathbb{R}^2 \mapsto \mathbb{C}$ is the complex pulse amplitude, $\beta_2 \in \mathbb{R}$ is the second order dispersion, and $\gamma \in \mathbb{R}$ is the coefficient of nonlinearity. In practice, (1) lacks a few key terms, thus, it is often generalized by adding amplification, loss, and occasionally higher order terms. This gives the generalized nonlinear Schrödinger equation (GNLSE) [2, 5, 11, 28, 29, 39],

$$\frac{\partial A}{\partial z} = -i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2} + i\gamma |A|^2 A + \frac{1}{2}g(A)A - \alpha A, \quad (2)$$

where $g(A)$ is an amplifying term due to the gain, and $\alpha \in \mathbb{R}$ is the loss due to scattering and absorption.

The GNLSE has many applications in nonlinear optics and fibre optic communications, however, in the context of lasers we typically also add a modulation term. This yields the master equation of mode-locking [13–16, 18, 20, 35, 37],

$$\frac{\partial A}{\partial z} = -i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2} + i\gamma |A|^2 A + \frac{1}{2}g(A)A - \alpha A - M(T), \quad (3)$$

where $M(T)$ is the modulation function. The solutions of three simplifications of (3) have been investigated:

- Omitting both modulation and nonlinearity [13, 15, 16].
- Omitting only modulation [19, 37].
- Omitting only nonlinearity [7, 13, 14, 17, 18, 20, 35, 37].

For a more comprehensive history see [18].

2.1 Discrete Component Models

While the derivation of (3) is sound mathematically, it is not representative of what happens within the laser cavity. The issue with (3) is that it has been assumed each process affects the pulse continuously within the cavity; for example, the pulse is amplified whether or not it is in the Erbium-doped fibre. As highlighted by Fig. 1, this is a rather poor assumption. Within the cavity each effect is localized to its corresponding component: almost all of the dispersion happens within the CFBG [1], the pulse is only amplified within the Erbium-doped fibre, etc. Thus, a better model is one where (3) is broken down into the individual components giving the effect of each ‘block’ of the cavity. Each of the blocks can then be functionally composed together to give an iterative map for the effect of one circuit around the cavity. This transforms the differential equation into an algebraic equation.

Such a method was first proposed in 1955 by Cutler [9] while analyzing a microwave regenerative pulse generator. This method was adapted for mode-locked lasers in 1969 by Siegman and Kuizenga [21, 30]. Kuizenga and Siegman also had success experimentally validating their model [22, 23]. The effects of the nonlinearity would not be considered until Martinez, Fork, and Gordon [26, 27] tried modelling passively mode-locked lasers. This issue has recently been readdressed by Burgoyne

[7] in the literature for tuneable lasers. In these models the effect of each component is described by a transfer function.

These discrete component models differ from the often used split-step Fourier method (see [2, 33]). The split-step Fourier method is a numerical technique used for solving nonlinear partial differential equations, such as (1). The method considers the linear and nonlinear terms separately and has a half integration step for both parts—in a manner similar to the leap-frog algorithm.³ Therefore, in the case of the NLSE, the dispersion and nonlinearity are still treated as continuous processes. However, in discrete component models the entire effect of each component is computed at once and the output of one component becomes the input for the following component, instead of alternating between the components in small integration steps. In this way, discrete component models are able to account for the geometry of a laser cavity, and indeed altering the permutation of the components gives rise to different dynamics.

Despite the development of discrete component models, several short-comings exist. The clearest is that none of these models have contained every block—either the nonlinearity or the modulation have been omitted. In the framework of tuneable lasers, each component plays a crucial role and the tuneable laser will not function correctly without the inclusion of all the components. Another key drawback is that the functional operations of some of the components used in their models are phenomenological. While these functions are chosen based on the observed output, they are not necessarily consistent with the underlying physics. Finally, none of these previous models have been able to exhibit a phenomenon called *modulation instability* in which the self-phase modulation of the pulse becomes too strong, distorting and damaging the wave until it ultimately becomes unstable and unsustainable.

3 A New Model

Using the ideas presented in the previous section of the prior functional models [7, 9, 21–23, 26, 27, 30] we shall derive a new model from (2)—with the exception of modulation in which we consider the exact functional form to be determined by the laser operator.

3.1 Components

We shall determine the effect each component has on the pulse by solving (2) while only considering the dominant term within each section of the cavity, and neglecting the others.

³ Also known as velocity Verlet in molecular dynamics, the Störmer method in astronomy, and further names in other areas [12].

3.1.1 Gain

Within the Er-doped gain fibre, the gain term is dominant, and Eq. (2) reduces to

$$\frac{\partial A}{\partial z} = \frac{1}{2}g(A)A, \quad (4)$$

where $g(A)$ takes the form [5, 7, 13, 14, 16, 18–20, 28, 29, 31, 37, 39]

$$g(A) = \frac{g_0}{1 + E/E_{\text{sat}}}, \quad E = \int_{-\infty}^{\infty} |A|^2 dT, \quad (5)$$

where g_0 is a small signal gain, E is the energy of the pulse, and E_{sat} is the energy at which the gain begins to saturate. Without much difficulty this differential equation can be solved, and the effect on an incident pulse is

$$G(A; E) = \left(\frac{E_{\text{out}}}{E}\right)^{1/2} A = \left(\frac{E_{\text{sat}}}{E} W \left(\frac{E}{E_{\text{sat}}} e^{E/E_{\text{sat}}} e^{g_0 L_g}\right)\right)^{1/2} A, \quad (6)$$

where L_g is the length of the gain fibre.

3.1.2 Nonlinearity

The nonlinearity of the fibre arises from the parameter γ . In regions where this effect is dominant expression (2) becomes

$$\frac{\partial A}{\partial z} - i\gamma|A|^2 A = 0. \quad (7)$$

Using a similar method as with the gain, the effect of the nonlinearity can be shown to be

$$F(A) = A e^{i\gamma|A|^2 L_f}, \quad (8)$$

where L_f is the length of fibre.

3.1.3 Loss

Expression (2) leads to exponential decay due to the scattering and absorption of the fibre. However, a majority of the signal is removed from the cavity by the optical coupler. Combining these two effects yields

$$L(A) = (1 - R)e^{-\alpha L_T} A, \quad (9)$$

where R is the reflectivity of the output coupler, and L_T is the total length of the laser circuit, as the effect of the losses.⁴

3.1.4 Dispersion

Considering only the dispersive term of (2), one obtains

$$\frac{\partial A}{\partial z} = -i \frac{\beta_2}{2} \frac{\partial^2 A}{\partial T^2}. \quad (10)$$

The effect of dispersion is then given by the map

$$D(A) = \mathcal{F}^{-1} \left\{ e^{i\omega^2 L_D \beta_2 / 2} \mathcal{F} \{A\} \right\}, \quad (11)$$

where L_D is the characteristic length of the dispersive medium, and \mathcal{F} denotes the Fourier transform.

3.1.5 Modulation

In this model, the modulation is considered to be applied externally in which ever way the operator sees fit. For simplicity the representation is taken as the Gaussian

$$M(A) = e^{-T^2 / 2T_M^2} A, \quad (12)$$

where T_M is the characteristic width of the modulation.

3.2 Non-dimensionalization

The structure of each process of the laser can be better understood by re-scaling the time, energy, and amplitude. Nominal values for tuneable lasers are shown in Table 1. Knowing experimental durations and energies, the table suggests the convenient scalings:

$$T = T_M \tilde{T}, \quad E = E_{\text{sat}} \tilde{E}, \quad A = \left(\frac{E_{\text{sat}}}{T_M} \right)^{1/2} \tilde{A}. \quad (13)$$

Revisiting each process map shows each process has a characteristic non-dimensional parameter. The new mappings—after dropping the tildes—are

⁴ Depending on the layout of the laser cavity the loss may take the form $L(A) = R e^{-\alpha L_T} A$ instead.

Table 1 Range of variation of various parameters

Parameter	Symbol	Value	Sources
Absorption of fibre ^a	α	10^{-4} – 0.3 m^{-1}	[6, 29, 36, 37, 39]
Fibre dispersion	β_2^f	-50 – $50 \text{ ps}^2/\text{km}$	[1, 2, 7, 25, 28, 39]
Fibre nonlinearity	γ	0.001 – $0.01 \text{ W}^{-1}\text{m}^{-1}$	[2, 11, 37, 39]
Grating dispersion	$\beta_2^g L_D$	10 – 2000 ps^2	[1, 2, 7, 24]
Length of cavity	L_T	10 – 100 m	[6, 28, 35]
Length of fibre	L_f	0.15 – 1 m	[6]
Length of gain fibre	L_g	2 – 3 m	[7, 28, 29, 34, 39]
Modulation time	T_M	15 – 150 ps	[5–7]
Reflectivity of optical coupler	R	0.1 – 0.9	[6, 24, 28, 34, 35, 38]
Saturation energy	E_{sat}	10^3 – 10^4 pJ	[6, 37, 39]
Small signal gain	g_0	1 – 10 m^{-1}	[6, 39]

^aFibre loss is typically reported as $\sim 0.5 \text{ dB/km}$

$$\begin{aligned}
 G(A) &= (E^{-1} W (a E e^E))^{1/2} A, & F(A) &= A e^{i b |A|^2}, & L(A) &= h A, \\
 D(A) &= \mathcal{F}^{-1} \left\{ e^{i s^2 \omega^2} \mathcal{F} \{A\} \right\}, & M(A) &= e^{-T^2/2} A,
 \end{aligned} \tag{14}$$

with the four dimensionless parameters, as defined by the values in Table 1,

$$\begin{aligned}
 a &= e^{g_0 L_g} \sim 8 \times 10^3, & h &= (1 - R) e^{-\alpha L_T} \sim 0.04, \\
 b &= \gamma L_f \frac{E_{\text{sat}}}{T_M} \sim 1, & s &= \sqrt{\frac{\beta_2 L_D}{2 T_M^2}} \sim 0.2,
 \end{aligned} \tag{15}$$

which characterize the behaviour of the laser. Notice that the modulation is only characterized by T_M , and each other process has its own independent non-dimensional parameter.

3.3 Combining the Effects of Each Block of the Model

In this model the pulse is iteratively passed through each process, the order of which must now be considered. We are most interested in the output of the laser cavity, and so we shall start with the loss component. Next, the pulse is passed through the CFBG, as well as the modulator. Finally, the pulse travels through the gain fibre to be amplified, and then we consider the effect of the nonlinearity since this is the region where the power is maximal. Note that in general the functional operators of the components do not commute, and therefore, the order of the components is indeed important—in contrast to previous models. This is especially the case for dispersion as realized through the Fourier transform. Functionally, this can be denoted as

$$\mathcal{L}(A) = F(G(M(D(L(A))))), \quad (16)$$

where \mathcal{L} denotes one loop of the laser. The pulse after one complete circuit of the laser cavity is then passed back in to restart the process. A steady solution to this model is one in which the envelope and chirp are unchanged after traversing every component in the cavity—we are uninterested in the phase. That is, such that $\mathcal{L}(A) = Ae^{i\phi}$ —for some $\phi \in \mathbb{R}$.

4 Conclusion

Within this paper we developed a nonlinear model for tuneable lasers. In order to better represent the underlying physics within the laser cavity, the nonlinear Schrödinger equation was reduced to simpler differential equations for each component of the laser. This led to a functional map that defines the effect of each component on a particular input pulse. These processes were then composed together to give an iterative mapping of the whole laser cavity. In a future publication, we shall show the results obtained by this iterative mapping as well as discuss the dynamics exhibited by this model—including modulation instability—to predict the conditions under which the pulse is stable and sustainable.

References

1. Agrawal, G.: *Fiber-Optic Communication Systems*, 3rd edn. Wiley, Inc. (2002)
2. Agrawal, G.: *Nonlinear Fiber Optics*, 5th edn. Academic Press (2013)
3. Al-Azzawi, A.: *Fiber Optics: Principles and Advanced Practices*, 2nd edn. CRC Press (2017)
4. Becker, P.C., Olsson, N.A., Simpson, J.R.: *Erbium-Doped Fiber Amplifiers Fundamentals and Technology*, 1st edn. Academic Press (1999)
5. Bohun, C.S., Cher, Y., Cummings, L.J., Howell, P., Mitre, T., Monasse, L., Mueller, J., Rouillon, S.: Modelling and specifying dispersive laser cavities. In: *Sixth Montréal Industrial Problem Solving Workshop*, pp. 11–25 (2015)
6. Burgoyne, B.: Private Communication (2018)
7. Burgoyne, B., Dupuis, A., Villeneuve, A.: An experimentally validated discrete model for dispersion-tuned actively mode-locked lasers. *IEEE J. Sel. Top. Quantum Electron.* **20**(5), 390–398 (2014). <https://doi.org/10.1109/JSTQE.2014.2303794>
8. Burgoyne, B., Villeneuve, A.: Programmable lasers: design and applications. In: *Proceedings of the SPIE*, vol. 7580 (2010). <https://doi.org/10.1117/12.841277>
9. Cutler, C.C.: The regenerative pulse generator. In: *Proceedings of the IRE*. IEEE (1955). <https://doi.org/10.1109/JRPROC.1955.278070>
10. Ferreira, M.F.S.: *Nonlinear Effects in Optical Fibers*. Wiley, Inc. (2011)
11. Finot, C., Kibler, B., Provost, L., Wabnitz, S.: Beneficial impact of wave-breaking for coherent continuum formation in normally dispersive nonlinear fibers. *J. Opt. Soc. Am. B* **25**(11), 1938–1948 (2008). <https://doi.org/10.1364/JOSAB.25.001938>
12. Hairer, E., Lubich, C., Wanner, G.: *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd edn. No. 31 in Springer Series in Computational Mathematics. Springer, Berlin; New York (2006)

13. Haus, H.A.: A theory of forced mode locking. *IEEE J. Quantum Electron.* **11**(7), 323–330 (1975). <https://doi.org/10.1109/JQE.1975.1068636>
14. Haus, H.A.: *Waves and Fields in Optoelectronics*. Prentice-Hall, Inc. (1984)
15. Haus, H.A.: Laser mode locking with addition of nonlinear index. *IEEE J. Quantum Electron.* **22**(2), 325–331 (1986). <https://doi.org/10.1109/JQE.1986.1072944>
16. Haus, H.A.: Analytic theory of additive pulse and Kerr lens mode locking. *IEEE J. Quantum Electron.* **28**(10), 2086–2096 (1992). <https://doi.org/10.1109/3.159519>
17. Haus, H.A.: Theory of soliton stability in asynchronous modelocking. *J. Lightwave Technol.* **14**(4), 622–627 (1996). <https://doi.org/10.1109/50.491401>
18. Haus, H.A.: Mode-locking of lasers. *IEEE J. Sel. Top. Quantum Electron.* **6**(6), 1173–1185 (2000). <https://doi.org/10.1109/2944.902165>
19. Haus, H.A., Fujimoto, J.G., Ippen, E.P.: Structures for additive pulse mode locking. *J. Opt. Soc. Am. B* **8**(10), 2068–2076 (1991). <https://doi.org/10.1364/JOSAB.8.002068>
20. Kärtner, F.: *Lecture Notes in Ultrafast Optics*. Massachusetts Institute of Technology: MIT OpenCourseWare (Online) (2005)
21. Kuizenga, D.J., Siegman, A.E.: FM and AM mode locking of the homogeneous laser—part I: theory. *IEEE J. Quantum Electron.* **6**(11), 694–708 (1970). <https://doi.org/10.1109/JQE.1970.1076343>
22. Kuizenga, D.J., Siegman, A.E.: FM and AM mode locking of the homogeneous laser—part II: experimental results in a Nd:YAG laser with internal FM modulation. *IEEE J. Quantum Electron.* **6**(11), 709–715 (1970). <https://doi.org/10.1109/JQE.1970.1076344>
23. Kuizenga, D.J., Siegman, A.E.: FM-laser operation of the Nd:YAG laser. *IEEE J. Quantum Electron.* **6**(11), 673–677 (1970). <https://doi.org/10.1109/JQE.1970.1076348>
24. Li, S., Chan, K.T.: Electrical wavelength tunable and multiwavelength actively mode-locked fiber ring laser. *Appl. Phys. Lett.* **72**(16), 1954–1956 (1998). <https://doi.org/10.1063/1.121263>
25. Litchinitser, N.M., Eggleton, B.J., Patterson, D.B.: Fiber Bragg gratings for dispersion compensation in transmission: theoretical model and design criteria for nearly ideal pulse recompression. *J. Lightwave Technol.* **15**(8), 1303–1313 (1997). <https://doi.org/10.1109/50.618327>
26. Martinez, O.E., Fork, R.L., Gordon, J.P.: Theory of passively mode-locked lasers including self-phase modulation and group-velocity dispersion. *Opt. Lett.* **9**(5), 156–158 (1984). <https://doi.org/10.1364/OL.9.000156>
27. Martinez, O.E., Fork, R.L., Gordon, J.P.: Theory of passively mode-locked lasers for the case of a nonlinear complex-propagation coefficient. *J. Opt. Soc. Am. B* **2**(5), 753–760 (1985). <https://doi.org/10.1364/JOSAB.2.000753>
28. Peng, J., Luo, H., Zhan, L.: In-cavity soliton self-frequency shift ultrafast fiber lasers. *Opt. Lett.* **43**(24), 5913–5916 (2018). <https://doi.org/10.1364/OL.43.005913>. <http://ol.osa.org/abstract.cfm?URI=ol-43-24-5913>
29. Shtyrina, O.V., Ivanenko, A.V., Yarutkina, I.A., Kemmer, A.V., Skidin, A.S., Kobtsev, S.M., Fedoruk, M.P.: Experimental measurement and analytical estimation of the signal gain in an Er-doped fiber. *J. Opt. Soc. Am. B* **34**(2), 227–231 (2017). <https://doi.org/10.1364/JOSAB.34.000227>
30. Siegman, A.E., Kuizenga, D.J.: Simple analytic expressions for AM and FM modelocked pulses in homogenous lasers. *Appl. Phys. Lett.* **6**, 181–182 (1969). <https://doi.org/10.1063/1.1652765>
31. Silfvast, W.T.: *Laser Fundamentals*, 2nd edn. Cambridge University Press (2004)
32. Starodoumov, A.N.: Optical fibers and accessories. In: Malacara-Hernández, D., Thompson, B.J. (eds.) *Advanced Optical Instruments and Techniques. Handbook of Optical Engineering*, vol. 2, 2nd edn., pp. 633–676. CRC Press (2018). Ch. 18. 2018
33. Taha, T.R., Ablowitz, M.J.: Analytical and numerical aspects of certain nonlinear evolution equations. II. Numerical, nonlinear Schrödinger equation. *J. Comput. Phys.* **55**(2), 203–230 (1984). [https://doi.org/10.1016/0021-9991\(84\)90003-2](https://doi.org/10.1016/0021-9991(84)90003-2)
34. Tamura, K., Ippen, E.P., Haus, H.A., Nelson, L.E.: 77-fs pulse generation from a stretched-pulse mode-locked all-fiber ring laser. *Opt. Lett.* **18**(13), 1080–1082 (1993). <https://doi.org/10.1364/OL.18.001080>

35. Tamura, K., Nakazawa, M.: Dispersion-tuned harmonically mode-locked fiber ring laser for self-synchronization to an external clock. *Opt. Lett.* **21**(24), 1984–1986 (1996). <https://doi.org/10.1364/OL.21.001984>
36. Tomlinson, W.J., Stolen, R.H., Johnson, A.M.: Optical wave breaking of pulses in nonlinear optical fibers. *Opt. Lett.* **10**(9), 457–459 (1985). <https://doi.org/10.1364/OL.10.000457>
37. Usechak, N.G., Agrawal, G.P.: Rate-equation approach for frequency-modulation mode locking using the moment method. *J. Opt. Soc. Am. B* **22**(12), 2570–2580 (2005). <https://doi.org/10.1364/JOSAB.22.002570>
38. Yamashita, S., Nakazaki, Y., Konishi, R., Kusakari, O.: Wide and fast wavelength-swept fiber laser based on dispersion tuning for dynamic sensing. *J. Sens.* **2009** (2009). <https://doi.org/10.1155/2009/572835>
39. Yarutkina, I., Shtyrina, O., Fedoruk, M., Turitsyn, S.: Numerical modeling of fiber lasers with long and ultra-long ring cavity. *Opt. Express* **21**(10), 12942–12950 (2013). <https://doi.org/10.1364/OE.21.012942>

Algebraic Structure and Complexity of Bootstrap Percolation with External Inputs



S. Pal and Chrystopher L. Nehaniv

Abstract In this paper a modification of the standard Bootstrap Percolation model is introduced. In our modification a discrete time update rule is constructed that allows for non-monotonicity—unlike its classical counterpart. External inputs to drive the system into desirable states are also included in the model. The algebraic structure and complexity properties of the system are inferred by studying the system’s holonomy decomposition. We introduce methods of inferring the pools of reversibility for the system. Dependence of system complexity on process parameters is presented and discussed.

Keywords Bootstrap percolation · Discrete-event dynamical systems · Transformation semigroups · Complexity analysis · Holonomy decomposition

1 Introduction

Bootstrap percolation is a process studied in statistical mechanics where cells in a lattice or any other space (for example, nodes in a random graph) exist in binary states—active or inactive. Given any initial configuration of states, the states of the nodes evolve with discrete time based on some predefined update rule. Bootstrap percolation is an example of a boolean network. The most popular update rule of bootstrap percolation is parameterized with a threshold k as follows:

$$x^i(t+1) = \begin{cases} x^i(t) & \text{if } x^i(t) = 1 \\ 1 & \text{if } x^i(t) = 0 \text{ and } \sum_{j \in N(i)} x^j(t) \geq k \\ 0 & \text{if } x^i(t) = 0 \text{ and } \sum_{j \in N(i)} x^j(t) < k, \end{cases}$$

S. Pal (✉)

Department of Applied Mathematics, University of Waterloo, Waterloo, Canada

e-mail: s24pal@uwaterloo.ca

C. L. Nehaniv

Faculty of Engineering, University of Waterloo, Waterloo, Canada

e-mail: cnehaniv@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_38

where $N(i)$ is the set of neighbours of node i and $x^i(t)$ is the boolean state of the node i at discrete time t : $x^i(t) = 1$ representing active state and $x^i(t) = 0$ representing the inactive state for node i at time t . The standard interpretation of the states may be inverted in some cases causing no loss of generality. There can be other update rules of a bootstrap percolation process like in [7] but all variants have the same property of being homogeneous and local. The standard model of bootstrap percolation is monotone with evolution of discrete time. Some variants of the bootstrap percolation, with the introduction of excitatory nodes and inhibitory nodes [7], display non-monotone behavior. The model of bootstrap percolation with inhibition has been useful in studying the phenomenon of input normalization in neurons [1]. In the next section a modification to the standard bootstrap percolation is introduced that also exhibits a non-monotonic behavior. Bootstrap percolation models have also been used to study impacts of external perturbations to models of weighted trade networks [8]. In [8], each node—a nation state—is modelled to be in a binary state. A node in an active state represents that the corresponding nation state is in a normal state, i.e. it has imports and exports in a stable level. An abnormal or inactive state for a nation indicates that it is facing a trading disaster and it is leading towards a volatile economy. Since an abnormal country can influence the state of its neighbours by potentially turning them into abnormal, bootstrap percolation is used for modeling the cascading reaction of any initial perturbation. Despite being an interesting model of the spread of trading and economic disasters in a trade network, certain major assumptions of the model make it less flexible to real settings. The model does not allow nodes going from inactive to active states through a local process in the same way they go from active to inactive. Moreover the model, in general does not allow independent relapse or recovery of nodes—i.e without the assistance of local homogenous update rule. Even though our modified bootstrap percolation model was motivated by drawbacks and inadequacies in the fault propagation model in trade-networks, one can use the structural framework of our model to represent other non-monotone local-homogenous processes.

2 The Modified Bootstrap Percolation Model with External Inputs

The new model description can be broken down into two independent parts:

1. The non-monotone bootstrap percolation process.
2. External inputs forcing certain network states to certain target network states.

The second part of the model has a superiority over the first part. After the formal description of these parts it will be clear as to how their hierarchy is defined

The non-Monotone Bootstrap Percolation Process:

This part of the model corresponds to the discrete step update rule. As opposed to the standard bootstrap percolation model, this modification has two process parameters k_1, k_2 instead of a single one k . The process is defined as:

$$x^i(t+1) = \begin{cases} 1, & \text{if } x^i(t) = 0 \text{ and } \sum_{j \in N(i)} x^j(t) \geq k_1 \\ 0, & \text{if } x^i(t) = 0 \text{ and } \sum_{j \in N(i)} x^j(t) < k_1 \\ 0, & \text{if } x^i(t) = 1 \text{ and } \sum_{j \in N(i)} x^j(t) \leq |N(i)| - k_2 \\ 1, & \text{if } x^i(t) = 1 \text{ and } \sum_{j \in N(i)} x^j(t) > |N(i)| - k_2 \end{cases}$$

In brief, the above process can be described textually as: (a) If a node is inactive and k_1 or more of its neighbours are active, it will turn active. (b) If a node is active and k_2 or more neighbours are inactive it will become inactive. Otherwise the state of nodes remains unchanged.

Note that now, with these changes we have an identical jump back rule from active to inactive like jumping from inactive to active. Unless there are any external inputs (the second part of the model), these update rules are obeyed at every discrete time step to evaluate the state of the network at the next time step. This part forms the local-homogenous update rule for our model.

External Inputs: Only the above modification does not allow nodes to have independent transitions to active or inactive states, i.e without the assistance of the homogenous update rule. In order to account for that, external forced inputs are introduced. Before defining the external forced inputs let's define what is meant by state of the network. Since a node can exist in two states and there are N nodes in the network, there can be 2^N possibilities of network state. Network state at discrete time step t is defined as $X(t) = (x^1(t), x^2(t), \dots, x^N(t))$. The network state is an element from $\{0, 1\}^N$. For convenience, each network state will be encoded with its decimal equivalent added with 1. So, for example the state $(1, 0, 0, \dots, 0)$ represents that only the node labelled 1 is active (1) and rest are inactive (0). If there were 4 nodes in the network the state $(1, 0, 0, 0)$ would have been encoded as 9. The 2^N states of a network are represented as X^1, X^2, \dots, X^{2^N} .

The external inputs are analogous to constant maps except for a crucial distinction: that some states of the network are non-forceable. This brings us to the third parameter of our model which is the set of states of the network which cannot be forced into a target set by any constant map. Formally, this part of the model can be expressed as:

For a network of size N the set of all possible external inputs is:

$$S_{\text{all}} = \{s_1, s_2, \dots, s_{2^N}\}$$

Let's define a one to one function F as: $F : s_i \mapsto X^i$. If S is defined as the set of non-forceable states of the network, then if at time step t there is an external input s_k to the process,

$$X(t + 1) = F(s_k) \text{ iff } X(t) \notin S$$

If $X(t) \in S$ then $X(t + 1)$ is determined by the local update rule of the process. At any time step t the model does not permit more than one external input to the process.

This part of the model has a hierarchy over the previous part. If there is any external input to the process, the external input part takes precedence over the local homogenous update rule.

3 Holonomy Decomposition and Modeling the Process as a Transformation Semigroup

In this paper the complexity and behaviour of such a modified model is analyzed by observing its algebraic structure. There are some obvious questions this paper aims to investigate. Since the process is non-monotone it is of interest to investigate whether a particular instance (a set of parameters k_1, k_2 and S) of this model falls into cycles or dies down into a single state. It may also be of interest to know whether it is possible to trigger the system into cycle—if yes, what is the fastest possible way of achieving that. Intuitively, the complexity of the process increases if the cardinality of S increases. In this section two measures are defined to quantify the algebraic complexity of the process. We will be using the method of computational Holonomy Decomposition of Transformation Actions to analyze this model [5]. The computational holonomy decomposition of any discrete-time, discrete state process requires the process to be represented as a transformation action. Before defining a transformation action representation the modified Bootstrap percolation model let’s look at holonomy decomposition and the advantages of using it to analyze our system.

A transformation action (also called a transformation semigroup) is defined as a set of functions that maps a set to itself and is closed under function composition. It is represented as (A, S) where S is the set of functions acting on the set A . (This S is not to be confused with the set of non-forceable states as defined in the model above.) Let $a \cdot s$ denote resulting state $s(a)$ in A resulting from applying transformation $s \in S$ to state $a \in A$. Similarly for $P \subseteq A$, $P \cdot s = \{a \cdot s \in A \mid a \in P\}$. The *extended image set* I^* of the transformation action (A, S) is defined as:

$$I^* = \{A \cdot s \mid s \in S\} \cup \{A\} \cup \{\{a\} \mid a \in A\},$$

A reflexive and transitive relation called the *subduction relation* is defined on I^* as:

$$P \leq_s Q \iff P = Q \text{ or } \exists s \in S \text{ such that } P \subseteq Q \cdot s \text{ for } P, Q \in I^*$$

Height of a singleton member Q in I^* is defined to be 0. For any other member $Q \in I^*$, the height is i is defined as the length of the longest strict subduction chain to Q in I^* that ends in a singleton, the Q ’s holonomy group (B_Q, H_Q) can be defined

as is done in [3]. A permutation reset semigroup $(\Gamma_i, \overline{\Theta}_i)$ for a height $i, i = 1, \dots, h$ is defined as the direct product of holonomy groups for subduction equivalence class representatives Q in I^* that have height i , augmented with the constant maps.

The *holonomy decomposition theorem* [3] states that any finite transformation semigroup *divides* or is *emulated* by a wreath product of its holonomy permutation reset transformation semigroups. Since holonomy decomposition is used to study the modified bootstrap percolation—the process is modeled as a transformation semigroup as follows:

$$(X = \{0, 1\}^N, \langle (S_{\text{all}} \setminus \{F^{-1}(s) | s \in S\}) \cup \{t\} \rangle) \tag{1}$$

Note that here S is the set of non-forceable states and S_{all} is the set of all external inputs. The angular brackets indicate that the semigroup acting on the state set X is generated by association of elements of the set inside the brackets. The actions of the transformation action is defined as:

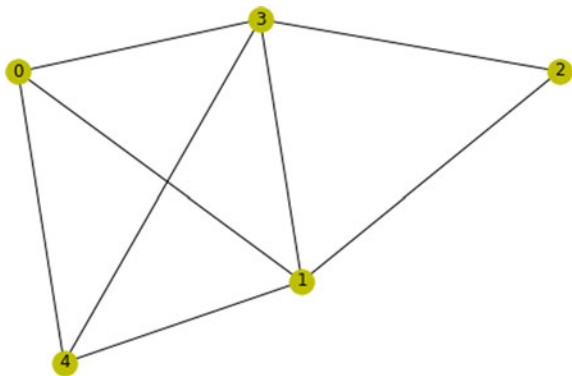
$$x \cdot t = y \implies X(t + 1) = y \text{ if } X(t) = x, \quad x, y \in X \text{ and no external input on system.}$$

$$\text{and } x \cdot s_i = F(s_i) \quad \text{if } x \notin S$$

$$x \cdot s_i = x \cdot t \quad \text{if } x \in S$$

For every parameter set k_1, k_2, S of the model for a given graph, a transformation semigroup can be written for the process using Eq. 1. For the remainder of this paper we show the results for the graph indicated in Fig. 1.

Fig. 1 Random graph with 5 nodes chosen for this study. A total of 32 states starting from 1 till 32 is possible for this network, for all the different boolean labellings of the nodes. Numbers inside the node indicate their index in the network statelabel



4 Methods, Results and Discussion

All the analysis of holonomy decomposition, algebraic structure, hierarchy and complexity were carried out using the computational tools of GAP (Groups, Algorithm, Programming)—a system for computational discrete algebra [12] and one of its packages: SgpDec (package for semigroup decomposition) [2]. In this paper, using GAP/SgpDec we study the holonomy decomposition of finite transformation actions [5, 6] to gain insights about the discrete system dynamics. Another concept that will be frequently used in this study is algebraic complexity of a process. We use two measures of complexity: (a) Krohn-Rhodes (KR) Complexity Measure (b) Aperiodic Complexity Measure. The KR complexity measure is a unique maximal hierarchical complexity measure satisfying the complexity axioms defined in [11]. It is defined as the smallest number of permutation levels needed in any Krohn-Rhodes decomposition [9]. In this paper we employ a computable upper bound on KR complexity of a transformation semigroup (TS) given as the total number of levels of the holonomy decomposition of TS with groups in them. We also define the *height complexity* of a transformation semigroup as the smallest number of levels needed in the holonomy decomposition of the transformation semigroup. It is to be noted that all the complexity measures reported in this paper are upper bounds of the actual KR complexity measure as GAP/SgpDec is never always guaranteed to produce a shortest Krohn-Rhodes decomposition for a transformation semigroup.

Using SgpDec it is possible to observe and study the skeleton of a transformation semigroup [4]. A skeleton of a transformation action is a pre-ordered structure on I^* which encodes information about the different ways of traversing from the entire state set to singletons of the system. It is possible to evaluate the upper bound of the number of consecutive irreversible transformations required (from any initialization) to reach a death state (or terminal generalized limit cycle, i.e., permutation group) for the system from the skeleton of a transformation action.

Since there are a lot of parameter combinations possible for our model, we focus on three scenarios of the process model which are somewhat realistic in the context of trade networks to demonstrate how holonomy decomposition can be used to study the bootstrap percolation process. Using those parameter values we will observe how the use of holonomy decomposition helps to understand the model process in more depth.

We study the following three scenarios of the modified Bootstrap percolation model. For all the three scenarios the values of k_1 and k_2 range from 1 to 5:

1. **Scenario 1:** In this scenario the set S (the set of non-forceable states) is empty.
2. **Scenario 2:** In this scenario the set S is taken as $\{11, 19, 20, 25\}$.
3. **Scenario 3:** In this scenario the set S is taken as $\{8, 15, 20, 22, 29\}$.

The choices for the set S in the above scenarios are not random. Their construction is dependent on the subject graph that has been chosen for this study. In the second scenario the motivation was to bunch five states into the non-forceable set S which allow only two nodes to be active with the constraint that their cumulative degree

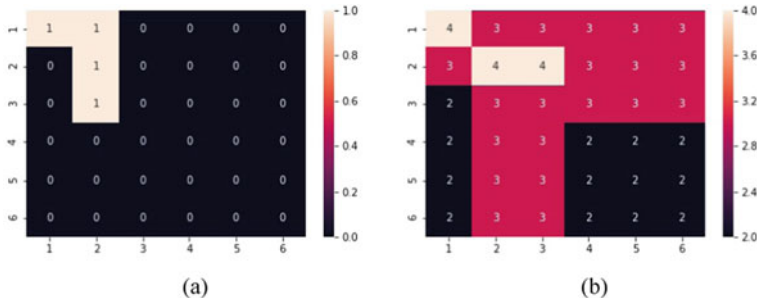


Fig. 2 Complexity analysis for Scenario 1. **a** Heat map for Krohn-Rhodes in scenario 1. **b** Heat map for height complexity in scenario 1. The horizontal axis of the heat map corresponds to the parameter k_1 . The vertical axis corresponds to the parameter k_2 . The numbers inside the cells represent the complexity value at the corresponding values of k_1 and k_2

does not exceed 6. Similarly in the third scenario the degree constraint was upper bounded to 10. The construction of these scenarios are realistic because in most applied cases, like the case of a trade network, making states non-forceable incurs certain costs.

Scenario 1 results: The upper bound for complexity measures for scenario 1 are presented in Fig. 2a, b with a heat map. For values of $(k_1, k_2) = (1, 1), (2, 1)$ and $(2, 3)$ the highest value for Krohn-Rhodes Complexity, i.e 1, is achieved. For $k_1 > 3$ and $k_2 > 3$ the Krohn-Rhodes complexity of the system is zero. This implies that in that subspace of the parameter space it is possible to construct an embedding of the model system by using only banks of flip-flops as there exists no pools of reversibility in the system in them. The system never falls into cycles and always dies into a state. The holonomy components of the system in scenario 1 as found by SgpDec for $k_1, k_2 = (1, 1)$ are: Level 1: 19, Level 2: 6, Level 3: 2, Level 4: $(7, C4)$

This encodes that the first level of the decomposition has 19 elements of extended image set being moved around by constant maps and identity. These transformation actions are called *identity resets*. We also see identity resets in levels 2 and 3. In the fourth level of the decomposition we have a permutation group $(7, C4)$ with a cyclic group $C4$ of order 4 acting on 7 elements, augmented with constant maps.

Although not presented in this report, it was found that the standard monotone bootstrap percolation model with constant maps (a simpler version than this one) has the simplest algebraic decompositions as it can be entirely built up with banks of flip flops. In other words the standard bootstrap percolation model with constant maps has Krohn-Rhodes complexity zero for all pairs of (k_1, k_2) and any other modification of that model can be viewed as being an added complexity to the fundamental model. The case of $k_1, k_2 = (1, 1)$ does however have cycles in it (indicated by permutation group in last level of decomposition). When the $(7, C4)$ holonomy group is investigated further using SgpDec, it was found that the states $\{6, 12, 10, 14\}$ are being moved by the generator t in a cycle, isomorphic to $C4$. In Fig. 3 we see the physical interpretation of this. The cycle in which the graph falls into has been

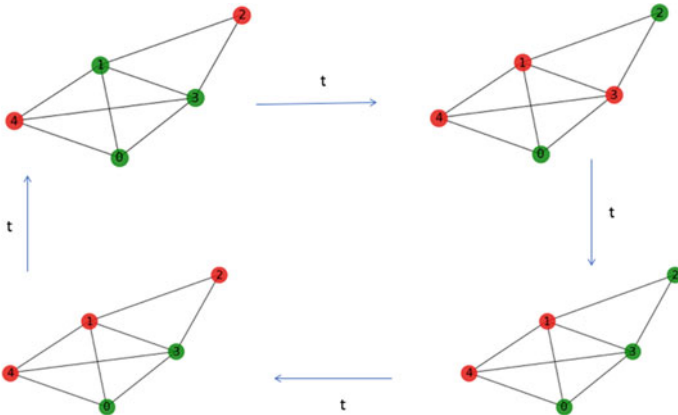


Fig. 3 The C4 cycle in last level of the holonomy decomposition of scenario 1 set at $(k_1, k_2) = (1, 1)$. Red nodes indicate inactive nodes and green nodes indicate active nodes

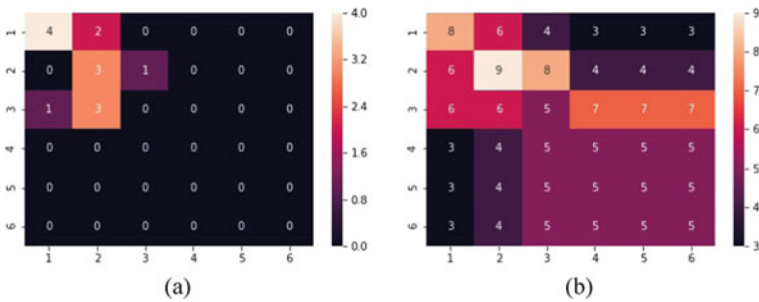


Fig. 4 Heat map for complexity measures over parameters k_1, k_2 for scenario 2. **a** KR Complexity upper-bound measure. **b** Height complexity measure. Highest KR complexity upper bound determined by SgpDec for this scenario is 4

identified exactly. If the system falls into any one of the states shown in Fig. 3 it will forever be stuck there (since we are in scenario 1 there is no forced exit from the loop as well).

Complexity increases in **Scenario 2** and **3** due to introduction of a non-forceable state as is observable in the heat maps of complexity in Figs. 4 and 5. Many more holonomy groups start to appear in the decomposition. Unlike scenario 1, in this scenario we have holonomy groups in levels in which there were none in the previous scenario. This implies that there are elements in the semigroup that move around sets of states together. To illustrate let’s take an example. At $(k_1, k_2) = (1, 1)$ for scenario 2, we find that the holonomy group $(5, C2)$ appears in the penultimate level of decomposition (the 7th). On investigating this group closely we see that the set of states $\{6, 8\}$ was being moved around by t^2 to $\{8, 10\}$ and back. This is shown in Fig. 6. In this case since an element of the transformation semigroup (here t^2 meaning two clock ticks) moves set of states around in cycles, it is less intuitive to understand

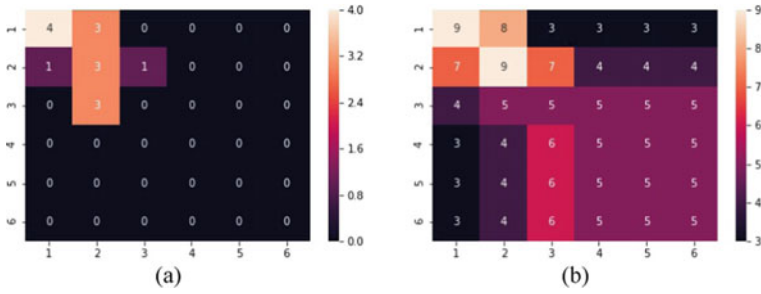


Fig. 5 Heat map for complexity measures over parameters k_1, k_2 for scenario 3. **a** KR complexity measure. **b** Height complexity measure

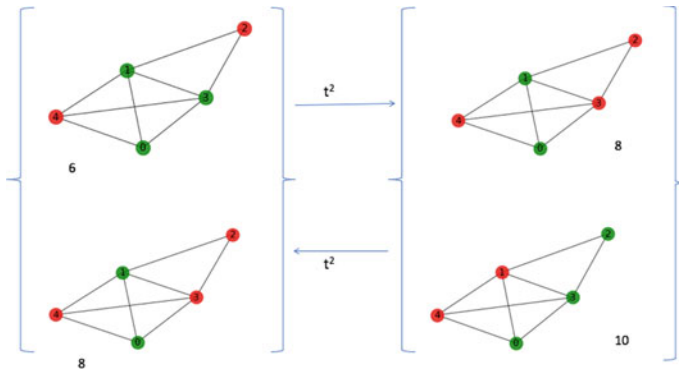


Fig. 6 The C2 cycle in the 7th level of the holonomy decomposition of Scenario 2 set at $(k_1, k_2) = (1, 1)$. Red nodes are inactive nodes and green nodes are active nodes. The set $\{6, 8\}$ is mapped to $\{8, 10\}$ and back by t^2

the actual cycles from the holonomy group itself. Looking at Natural Subsystems can help in having a better insight in these cases [10].

We also see that the highest KR complexity bound attained in Scenario 2 and 3 is 4. Scenario 3 is only a bit more complex than Scenario 2 in the height complexity measure. For some parameter tuples (k_1, k_2) the upper bound of the number of irreversible actions required on the original state set to bring system to a death state is higher for Scenario 3 when compared to Scenario 2 (and of course, scenario 1). Note that in Scenario 2 and 3, just like Scenario 1, there exist threshold values for k_1 and k_2 above which KR complexity measure is zero as no pools of reversibility can exist in those parameter spaces.

5 Conclusion

In this paper we introduced a modified non-monotone version of the classical bootstrap percolation that allowed external inputs and modelled it as an automaton in order to analyze its algebraic structure and corresponding complexity. Existence of cycles can be beneficial or detrimental depending on the system being studied. So their identification in a discrete event discrete time system process can be of interest. The complexity measures identified in this paper provide insight on that matter. A non-zero KR complexity measure for discrete systems indicates the existence of subsets of the state space that can be moved around in cycles by sequence of combination of external operations. We also discuss analysis by holonomy decomposition of our bootstrap percolation model and discuss how it reveals hidden structures and information about the system. The decomposition reveals that our system's hierarchical construction can be emulated by building blocks of cyclic groups, simple non-abelian groups and flip-flops. Apart from cycles, other interesting system properties are also identified for our system by studying metrics like height complexity of a holonomy decomposition. The non-monotone version of the Bootstrap percolation model under discussion shows higher complexity in structure than the classical Bootstrap Percolation model. Adding a non-forceable state set on top of that results in further increases in the complexity measures.

Acknowledgements This work was supported in part by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant, funding ref. RGPIN-2019-04669

References

1. Carandini, M., Heeger, D.J.: Erratum: normalization as a canonical neural computation (Nat. Rev. Neurosci. **13**(51–62) (2012)). Nat. Rev. Neurosci. **14**(2), 152 (2013)
2. Egri-Nagy, A., Nehaniv, C., Mitchell, J.D.: SgpDec—software package for Hierarchical Composition and Decomposition of Permutation Groups and Transformation Semigroups. <https://github.com/gap-system/sgpdec> (2015)
3. Egri-Nagy, A., Nehaniv, C.L.: Cycle structure in automata and the holonomy decomposition. Acta Cybern. **17**(2), 199–211 (2005)
4. Egri-Nagy, A., Nehaniv, C.L.: On the skeleton of a finite transformation semigroup. In: Annales Mathematicae et Informaticae, vol. 37, pp. 77–84 (2010)
5. Egri-Nagy, A., Nehaniv, C.L.: Computational holonomy decomposition of transformation semigroups. [arXiv:1508.06345](https://arxiv.org/abs/1508.06345) (2015)
6. Eilenberg, S.: Automata, Languages and Machines, vol. B. Academic Press (1976)
7. Einarsson, H., Lengler, J., Panagiotou, K., Mousset, F., Steger, A.: Bootstrap percolation with inhibition. [arXiv:1410.3291](https://arxiv.org/abs/1410.3291) (2014)
8. Fan, Y., Ren, S., Cai, H., Cui, X.: The state's role and position in international trade: a complex network perspective. Econ. Model. **39**, 71–81 (2014)
9. Krohn, K., Rhodes, J.: Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. Trans. Am. Math. Soc. **116**, 450–464 (1965)

10. Nehaniv, C.L., Rhodes, J., Egri-Nagy, A., Dini, P., Morris, E.R., Horváth, G., Karimi, F., Schreckling, D., Schilstra, M.J.: Symmetry structure in discrete models of biochemical systems: natural subsystems and the weak control hierarchy in a new model of computation driven by interactions. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **373**(2046), 20140223 (2015)
11. Nehaniv, C.L., Rhodes, J.L.: The evolution and understanding of hierarchical complexity in biology from an algebraic perspective. *Artif. Life* **6**(1), 45–67 (2000)
12. The GAP Group: GAP—Groups, Algorithms, and Programming, Version 4.11.0. <https://www.gap-system.org> (2020)

Simulations of Realistic Trombone Notes in the Time-Domain



Janelle Resch, Lilia Krivodonova, and John Vanderkooy

Abstract Time pressure waveforms associated with four musical notes produced at different volumes on a trombone were measured and then reproduced using a computational model. Special care was taken to accurately model the change in the trombone's cross-sectional area. An axisymmetric formulation of the compressible Euler equations was used and then numerically solved using the discontinuous Galerkin method. To evaluate the chosen model, the numerical solutions were compared against the measured data collected outside the bell. We found that accounting for the nonlinear behaviour for both high and low sound volumes yielded a good quantitative match between the computed and measured tones. For all four notes, once the sound pressure level drops 30 dB below the main peak, the computed pressure overestimated the measured spectral components.

Keywords Nonlinear acoustics · Compressible Euler equations · Discontinuous Galerkin method · Wave steepening · Trombone

1 Introduction

In this paper, we present results of axisymmetric simulations of sound propagation in a trombone where notes of different dynamic levels were considered. Modeling the state of a musical instrument during play, i.e., simulating realistic musical notes,

J. Resch (✉) · L. Krivodonova
Department of Applied Mathematics, University of Waterloo, 200 University Ave. W.,
Waterloo, ON N2L 3G1, Canada
e-mail: jresch@uwaterloo.ca

L. Krivodonova
e-mail: lgk@uwaterloo.ca

J. Vanderkooy
Department of Physics and Astronomy, University of Waterloo, 200 University Ave. W.,
Waterloo, ON N2L 3G1, Canada
e-mail: jv@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_39

is equivalent to recreating the timbre produced by the vibrating air-column [16]. If a note is played softly on a brass instrument, then its tonal character is typically expressed by the first few harmonics [4]. In acoustics, words such as ‘mellow’ or ‘dull’ have been used to characterize such timbres [7]. But as the playing dynamic or loudness of the note increases, the amplitude as well as the rate of change of the pressure disturbance entering the bore increases. This results in the sound quality becoming more ‘rich’ or ‘bright’ [3, 6, 8, 13]. This tonal character or ‘brassiness’ is the acoustic consequence of nonlinear wave propagation, i.e., the distortion of the waveform’s shape as it travels through the tubing of the instrument. In particular, the crest of the pressure wave will travel faster than the trough causing the wave to steepen. Such wave propagation is especially prevalent in the trombone due to its length and general shape [7]. For such strong nonlinear behaviour, a linearization cannot be applied to the equations of motion, particularly if a loud note is being modeled. This has been shown for instance in [9, 10, 12].

In this work, we further demonstrate that regardless of a note’s volume, realistic musical tones can be accurately simulated when the nonlinear motion is incorporated into the model. This is accomplished by measuring and then simulating the pressure waveform associated with a B_3^b played at mezzo-piano, a B_3^b and B_4^b played at forte, and a F_3 played at double-forte through a simplified trombone geometry. Since we previously showed [15] that the bends of the instrument do not greatly influence the wave propagation, the trombone can be thought of as a tube with axial symmetry. We exploited this symmetry and used an axisymmetric model where the change in the trombone’s radius was carefully reconstructed.

2 Computational Model

We write the general conservation law in a domain Ω as

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{0}, \quad \mathbf{x} \in \Omega, \quad t > 0, \quad (1a)$$

$$\mathbf{u} = \mathbf{u}^0, \quad t = 0, \quad (1b)$$

where $\mathbf{f}(\mathbf{u})$ is the flux function and the solution is $\mathbf{u}(\mathbf{x}, t) = (u_1, u_2, \dots, u_m)^t$, $(\mathbf{x}, t) \in \Omega \times [0, T]$. The solution $\mathbf{u}(\mathbf{x}, t)$ on each element is approximated by a vector function \mathbf{U}_j whose components are written as a linear combination of the orthogonal basis functions $\{\varphi_j\}$.

We model nonlinear sound wave propagation through a trombone using the compressible Euler equations in which we describe the flow as an inviscid, isentropic fluid. The equations of motion will be written using the 2D axisymmetric system (x, r) where r is the radial component and x is along the trombone axis. The solution is independent from the angular coordinate θ . To avoid the singularity at $r = 0$, the surface integral is not computed along the axis of symmetry and the vector of

conserved variables \mathbf{U} is multiplied by r . Taking ρ as the gas density, p as the internal pressure, E as the total energy, $\rho\mathbf{u} = (\rho u, \rho v)$ as the momenta in the axial and radial directions, the system in (x, r) coordinates can be written as

$$\frac{\partial[r\mathbf{U}]}{\partial t} + \frac{\partial[r\mathbf{F}(\mathbf{U})]}{\partial x} + \frac{\partial[r\mathbf{G}(\mathbf{U})]}{\partial r} = \mathbf{S}(\mathbf{U}), \quad (2)$$

where the flux vectors $\mathbf{F}(\mathbf{U})$, $\mathbf{G}(\mathbf{U})$, and the source term $\mathbf{S}(\mathbf{U})$ are defined as

$$\mathbf{F}(\mathbf{U}) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E + p) \end{bmatrix}, \quad \mathbf{G}(\mathbf{U}) = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E + p) \end{bmatrix}, \quad \mathbf{S}(\mathbf{U}) = \begin{bmatrix} 0 \\ 0 \\ p \\ 0 \end{bmatrix}. \quad (3)$$

The equation of state for an ideal gas connects E to the other variables and closes the system

$$E = r \left(\frac{P}{\gamma - 1} + \frac{\rho}{2} (u^2 + v^2) \right), \quad (4)$$

where the parameter γ is the specific heat ratio, which for air is $\gamma = 1.4$ [17].

2.1 Numerical Test Case

We simulated the time pressure waveform of the recorded B_3^b played at mezzo-piano (*mp*), a B_3^b and B_4^b played at forte (*f*), and a F_3 played at double-forte (*ff*) (see [15] for details). These pressure measurements were obtained by mounting a quarter-inch microphone on a Mendini MTB-L B^b tenor slide trombone approximately 4.7 cm from the beginning of the mouthpiece. This area of the trombone is known as the *mouthpiece-shank*. The wave profiles were prescribed as the boundary conditions on pressure at the inlet boundary of the computational trombone for the simulations presented in Sect. 3. We wrote an expression for the pressure by applying Fourier synthesis to one period of the recorded waveform. The series was truncated at the 31st term and written as

$$p = A_0 + 2 \sum_{i=1}^{30} A_i \cos(2\pi f_i t + \phi_i), \quad (5)$$

where f_i denotes an integer multiple of the fundamental frequency, A_i and ϕ_i is the amplitude and phase corresponding to each harmonic component, respectively, and A_0 is the term corresponding to the direct current.

Since the cross-sectional area near mouthpiece-shank only increases slightly for the initial 4.5 cm of tubing and then remains constant for approximately 159.5 cm, we locally related pressure and velocity at the mouthpiece boundary using the planar

expression derived from linear acoustic theory. This relation between pressure and velocity reproduced the measured mouthpiece pressure waveform accurately. Finally, the density was prescribed assuming the adiabatic relation between pressure and density from compressible flow theory [14]. In summary, the dimensionless boundary conditions at the mouthpiece of the computational trumpet are given by

$$\begin{cases} p = A_0 + \sum_{i=1}^{30} 2A_i \cos(2\pi f_i t + \phi_i), \\ \rho = \gamma p^{\frac{1}{\gamma}}, \\ u = \frac{p-p_0}{\rho_0 c}, \\ v = 0.0. \end{cases} \quad (6)$$

The computed pressure was sampled 17 cm outside the computational trombone. This corresponds to the position where another microphone was placed along the central axis of the real instrument. The microphones simultaneously recorded the mentioned notes so we could examine the evolution of the waveform as it traveled through the instrument. Comparing the experimental waveforms outside the bell with our numerical outputs allowed us to test the validity of our model.

2.2 Initial and Boundary Conditions

We modeled a trombone where the flare opens into an open domain and took the flow to initially be at rest. For all simulations, the flow (6) is introduced into the domain at the left vertical boundary of the bore which corresponds to the mouthpiece boundary. Along the far-field, pass-through boundary conditions were used in which the ghost state was prescribed to be the free flow state, i.e., the initial state. We experimentally determined the size of the computational domain so that reflections at the far-field would not influence the waveform solution. Reflective boundary conditions were prescribed (i.e., solid-wall boundary conditions) on the inner and outer walls of the computational instrument. At the ghost state, the normal velocity was taken to be the inner value with a negative sign. The density, pressure and tangential velocity were unchanged from the corresponding values inside the cell (see [5] for more details).

2.3 Computational Trombone Geometry

We now present the computational trombone that was used for our numerical simulations. The computational geometry describes the physical shape of the 2.87 m long trombone where the flow in the mouthpiece cup was not considered. A general diagram of the trombone mouthpiece and the positioning of the first microphone is shown in Fig. 1. The shaded region represents the beginning of our computational domain.

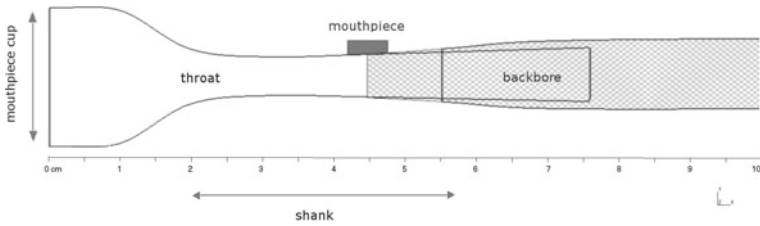
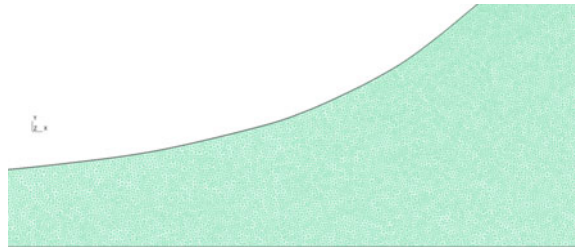


Fig. 1 A diagram of the first 10 cm of the trombone. The shaded region corresponds to the beginning of the computational domain where the left vertical wall at 4.7 cm is the mouthpiece boundary. The junction between the trombone tubing and the mouthpiece is located at 5.6 cm

Fig. 2 An example of the mesh generated using GMSH inside the computational trombone flare region



The first 4.5 cm of the trombone bore is conical (not including the mouthpiece cup), whereas the next 159.5 cm of tubing remains cylindrical (the first bend is within this region). Leading into the second bend however, the trombone tubing slightly increases in radius again. The second bend then immediately leads into the region of the rapidly expanding flare. Therefore, if the trombone were to be straightened out, between 164 to 247 cm, the bore is a conical shape whose radius increases by a factor of 1.59. To obtain a realistic flare shape, a photograph of the trombone bell was taken. The *grabit* software (Math Works Inc.) was then used to trace out the trombone flare by a set of points. Cubic splines were used to interpolate the bell and initial bore shape and lines were used for the cylindrical regions. We will refer to this computational geometry as $Geo_{Trombone}$. The corresponding mesh was obtained using the mesh generating software GMSH. In Fig. 2, a close up of the mesh inside the flare region is shown. Adaptive element sizes were used to accurately resolve the geometric features of the trombone. The final mesh had a total of 935,366 triangular shaped cells where the minimum inscribed radius was 75.4 μm .

3 Simulation Results

The 2D axisymmetric compressible Euler equations (3)–(4) were solved on $Geo_{Trombone}$ where the B_3^b at *mp*, the B_3^b at *f*, the B_4^b at *f*, and the F_3 at *ff* were generated at the inflow boundary. All four computed pressure waveforms were sampled 17 cm

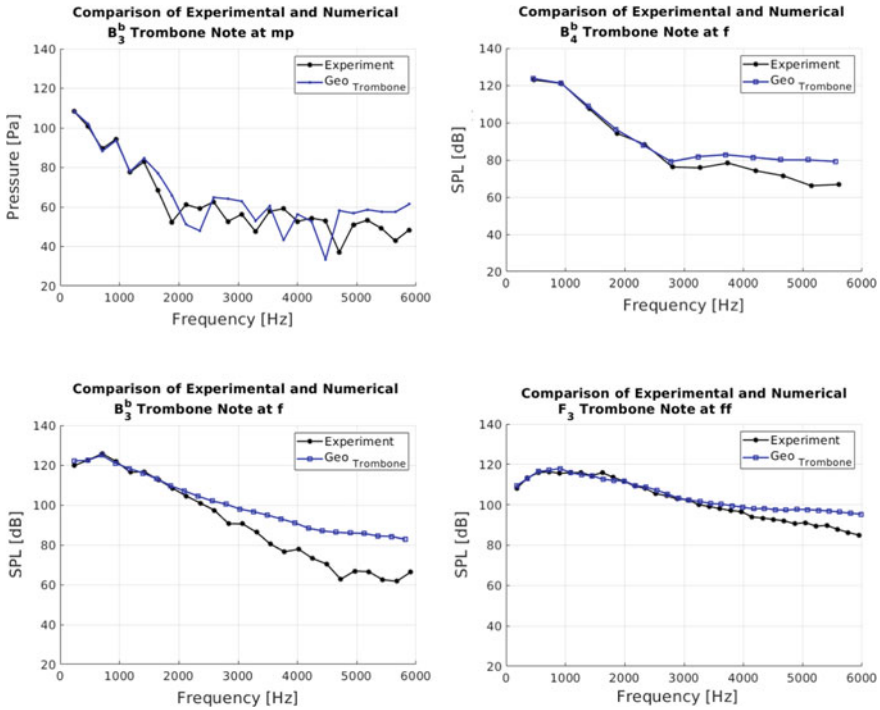


Fig. 3 A comparison between the experimental and computed frequency spectra of the B_3^b played at mp (top-left), B_4^b played at f (top-right), B_3^b played at f (bottom-left) and F_3 played at ff (bottom-right) simulated on $Geo_{Trombone}$

outside the bell and then compared against the measured data at the same position. The computed and measured spectral curves for these notes are plotted in Fig. 3.

For the mp trombone note, the sound pressure levels (SPLs) corresponding to the first six frequencies (i.e., components less 1500 Hz) match well with the experimental data. These components are most important since it is a softly played note, and the higher measured harmonics appear to contain mostly noise. We will therefore dismiss the noisy portion of the spectrum. Next we will evaluate whether brassy timbres, i.e., louder notes, could also be accurately reproduced. Examining the B_3^b/B_4^b played at f in Fig. 3, we found that the simulated notes match the measured data exceptionally well for all frequencies 2800 Hz. A lower, louder pitch—the F_3^b played at ff , was also simulated on $Geo_{Trombone}$ yielding even better results for frequencies up 4000 Hz. The relative differences in the SPLs between the computed and measured pitches are plotted in Fig. 4.

Although we have not presented the results in this paper, in [15] we examined the general importance of incorporating nonlinear effects (i.e., wave steepening) when simulating musical notes, especially when attempting to reproduce musical notes through the trombone when shock waves are produced. In particular, simulations of

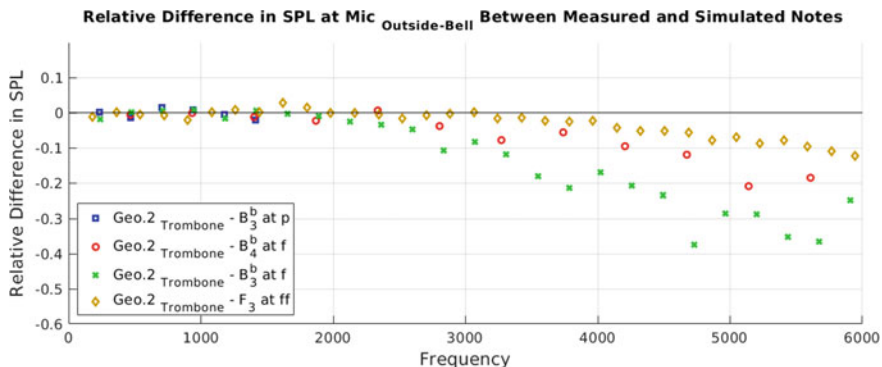


Fig. 4 Relative difference in SPLs between the measured and simulated trombone notes shown in Fig. 3

both finite-amplitude (nonlinear) and small-amplitude (linear) musical tones were simulated and their spectra were compared. It was found that when wave steepening was neglected, the numerical solutions greatly underestimated the amplitude of harmonic components larger 1500 and 500 Hz for the trumpet and trombone, respectively, when modeling a B_3^f played at forte.

4 Conclusion

In the literature, it is typical to consider six to ten harmonics when analyzing the timbre of f notes, [1, 2, 7, 11, 12]. By these standards, our proposed model is able to reproduce the brassiness of the mentioned notes rather well. Regardless of the playing dynamic level, deviations from the experimental data were observed mainly for the highest frequencies. In particular, for harmonics with the SPLs that are roughly 30 dB below the maximum SPL, the computed spectra overestimates the measured values where the discrepancy increases with frequency. Nonetheless, the lower and mid-frequencies of the trombone notes matched the experimental spectra very well. This makes it tempting to suggest that the observed variation is due to neglecting thermoviscous effects (since losses are more efficient for higher frequencies). Future work should attempt to incorporate such losses.

Acknowledgements This research was supported in part by the Alexander Graham Bell PGS-D grant 365873. We also acknowledge and thank the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Adachi, S., Sato, M.A.: Time-domain simulation of sound production in the brass instrument. *J. Acoust. Soc. Am.* **97**(6), 3850–3861 (1995)
2. Backus, J., Hundley, T.C.: Harmonic generation in the trumpet. *J. Acoust. Soc. Am.* **49**, 509–519 (1971)
3. Benade, A.H.: *Fundamentals of Musical Acoustics*. Dover Publications, New York (1900)
4. Campbell, M., Chick, J., Gilbert, J., Kemp, J., Myers, A., Newton, M.: Spectral enrichment in brass instruments due to nonlinear sound propagation; a comparison of measurements and predictions. In: *Proceedings of ISMA, Le Mans* (2014)
5. Flaherty, J.E., Krivodonova, L., Remacle, J.F., Shephard, M.S.: Some aspects of discontinuous Galerkin methods for hyperbolic conservation laws. *J. Finite Elem. Anal. Des.* **38**(10), 889–908 (2002)
6. Fletcher, N.H., Rossing, T.D.: *The Physics of Musical Instruments*. Springer Science & Business Media (2012)
7. Hirschberg, A.J., Gilbert, J., Msallam, R., Wijnands, A.P.J.: Shock waves in trombones. *J. Acoust. Soc. Am.* **99**(3), 1754–1758 (1995)
8. Logie, S.M.: Acoustical study of the playing characteristics of brass wind instruments. Ph.D. thesis, University of Edinburgh (2013)
9. Msallam, R., Dequidt, S., Caussé, R., Tassart, S.: Physical model of the trombone including nonlinear effects. Application to the sound synthesis of loud tones. *Acta Acust. United Acust.* **86**(4), 725–736 (2000)
10. Noreland, J.D., Udawalpola, M.R., Berggren, O.M.: A hybrid scheme for bore design optimization of a brass instrument. *J. Acoust. Soc. Am.* **128**(3), 1391–1400 (2010)
11. Petiot, J.F., Gilbert, J.: Comparison of trumpets' sounds played by a musician or simulated by physical modelling. *Acta Acust. United Acust.* **99**(4), 629–641 (2013)
12. Rendón, P.L., Narezo, D., Bustamante, F.O., Lopez, A.P.: Nonlinear progressive waves in a slide trombone resonator. *J. Acoust. Soc. Am.* **127**(2), 1096–1103 (2009)
13. Rendón, P.L., Ezeta, R., Pérez-López, A.: Nonlinear sound propagation in trumpets. *Acta Acust. United Acust.* **99**(4), 607–614 (2013)
14. Resch, J., Krivodonova, L., Vanderkooy, J.: A two-dimensional study of finite amplitude sound waves in a trumpet using the discontinuous Galerkin method. *J. Comput. Acoust.* **22**(3), 27 (2014)
15. Resch, J.: Physical modelling and associated acoustic behaviour of trumpets and trombones. Ph.D. thesis, University of Waterloo (2019)
16. Rocamora, M., Lopez, E., Jure, L.: Wind instruments synthesis toolbox for generation of music audio signals with labeled partials. In: *Proceedings of 2009 Brazilian Symposium on Computer Music*, vol. 2, pp. 2–4 (2009)
17. Warburton, T., Hesthaven, J.S.: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer, New York (2008)

The Impact of External Features on Prediction Accuracy in Short-Term Energy Forecasting



Maher Selim, Ryan Zhou, Wenying Feng, and Omar Alam

Abstract Accurate prediction of future electricity demand is important in the energy industry. Machine learning for time series prediction provides solutions for short term energy forecasting through a variety of algorithms, such as LSTM, SVR, Xgboost, and Facebook Prophet. However, many companies primarily rely on univariate time series algorithms, while numerous external data, e.g. weather data, are available as input features for energy forecasting. In this paper, we study the impact of external features on the performance of univariate and multivariate time series algorithms for Short-term Energy Forecasting using a standard benchmark energy data set. Quantitative comparisons on prediction accuracy measured by Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for the models are obtained. It is found that multivariate algorithms using external features outperform univariate algorithms, and that multivariate algorithms achieve reasonable accuracy even without using past step energy consumption as an input feature.

Keywords Energy forecasting · Facebook prophet · LSTM · Machine learning · Support vector regression

1 Introduction

Forecasting energy demand is critical for the energy industry as well as businesses in related sectors, such as banks and insurance companies. It has been estimated that a one-percent improvement in mean absolute percentage error (MAPE) can save \$300,000 annually for a utility company with a 1 GW peak load, and millions of dollars for larger ones [1]. An accurate forecast of upcoming energy consumption allows utility companies to plan and make decisions in real-time for all processes in their system and is a requirement to build automated smart energy grids [2, 3]. However, this problem has increasingly become more complex in recent years due

M. Selim · R. Zhou · W. Feng (✉) · O. Alam
Trent University, Peterborough, ON, Canada
e-mail: wfeng@trentu.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_40

to growing energy markets [1] and the introduction of renewable sources which are tightly coupled with external variables such as weather conditions. Despite the fact that energy forecasting is increasingly becoming a multivariate problem, many companies in the energy sector continue to use univariate time series algorithms, which only consider the usage history of electricity consumption.

Electricity forecasting using data-driven approaches, such as, machine learning is the subject of ongoing research [1]. A recent survey [1] shows that the percentages of machine learning algorithms investigated for Short-Term Energy Load Forecasting (STELF) are as follows: 4% decision trees, 24% statistical and other algorithms, 25% support vector machines (SVM), and 47% artificial neural networks (ANN) [1, 4, 5].

Research in energy forecasting focuses primarily on improvement of univariate models [6, 7]. Our contributions in this paper are twofold. First, we test representative models using algorithms from each of the above surveyed categories [1], and demonstrate that multivariate approaches consistently outperform the univariate model Facebook Prophet for energy forecasting. To this end, four computational models are adopted to our research purpose and tested: Long Short-Term Memory neural networks (LSTM) [8], Support Vector Regression (SVR) [9], Gradient Boosted Trees [10], and Facebook Prophet [11]. Prediction accuracy for all models is compared with Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Second, we show that even when the past energy load is excluded as a feature, the models are capable of predicting future load based on external features and even when the features are measured on a larger timescale than the target variable. These results are of interest to the energy industry as they demonstrate a simple and computationally light method to improve currently used models.

The remainder of this paper is organized as follows. Section 2 provides an overview of machine learning time series algorithms studied. Section 3 discusses the computational models and methodology implementation. Section 4 explains our experimental results and Sect. 5 concludes the paper.

2 Notable Multivariate and Univariate Machine Learning Time Series Algorithms

We first briefly discuss the mathematical background for notable time series algorithms including the multivariate algorithms LSTM [8], SVR [9], Gradient Boosted Trees via XGBoost [10], and the univariate Facebook Prophet package [11]. This forms the basis for our model development and implementation to be explained in Sect. 3.

Univariate and multivariate time series models: The univariate time series is a set of continuous observations for a single variable with constant time steps [12], univariate models aim to predict future values for that single variable based only on its past values:

$$\hat{x}_t = F(x_{t-1}, x_{t-2}, x_{t-3}, \dots)$$

where x_t represents the value of the target variable at time t , and F is the learned function. The multivariate time series is defined as observations of one or more variables and features often taken simultaneously and describes the interrelationships among the series [13–15]. Multivariate models use variables and features of time-series data to develop a model to forecast future prediction for the target variable:

$$\hat{x}_t = F(x_{t-1}, x_{t-2}, x_{t-3}, \dots, a_{t-1}^{(1)}, a_{t-2}^{(1)}, a_{t-3}^{(1)}, \dots, a_{t-1}^{(2)}, a_{t-2}^{(2)}, a_{t-3}^{(2)}, \dots)$$

where each $a^{(i)}$ represents the time series of an external feature. We also investigate models of the form

$$\hat{x}_t = F(a_{t-1}^{(1)}, a_{t-2}^{(1)}, a_{t-3}^{(1)}, \dots, a_{t-1}^{(2)}, a_{t-2}^{(2)}, a_{t-3}^{(2)}, \dots)$$

where past information about the target variable is unavailable.

Long Short-Term Memory (LSTM) neural networks: LSTM is a type of recurrent neural network architecture designed to extract long-term dependencies out of sequential data and avoid the vanishing gradient problem present in ordinary recurrent networks [16, 17]. These properties make LSTM the method of choice for longer time series and sequence prediction problems [18, 19]. LSTMs have been successfully applied to Short-Term Electricity Load Forecasting (STELF) modeling [2, 8]. There are several variations of the LSTM unit, but in this paper we use the standard architecture designed by Graves and Schmidhuber [16].

The key idea behind LSTM is to introduce a memory cell to the standard RNN architecture [2, 8]. This memory cell allows the LSTM module to retain information across many timesteps when needed [18, 19].

Support Vector Regression (SVR): Nonlinear support vector regression is an extension of the support vector machine (SVM) to regression problems [20]. The statistical learning theory for support vector regression is developed in [21]. Assuming that $D = \{x_i, y_i\}_{i=1}^n$ is a training dataset, where $x_i \in R^d$ are the system features and $y_i \in R$ is the main system output observations, the goal of ϵ -SVR is to find a function $f(x)$ that has no more than ϵ deviation from the observed output y_i for all training data.

This leads to the SVR optimization:

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i + \hat{\xi}_i) \\ \text{s.t.} \quad & y_i - w^T x_i - \xi_i \leq \epsilon \\ & -(y_i - w^T x_i) - \hat{\xi}_i \leq \epsilon \\ & \xi_i \geq 0, \hat{\xi}_i \geq 0, i = 1, \dots, n. \end{aligned}$$

where w is the learned weight vector, n denotes the number of samples, x_i is the i -th training instance, y_i is the training label, and ξ_i the distance between the bounds and predicted values outside the bounds. C is a parameter set by the user that controls the penalty imposed on observations outside the bounds, which helps to prevent overfitting. The SVR uses kernel functions to transform the data into a higher dimensional feature space to make it possible to perform the linear separation. In this paper, we use three different kernels with SVR namely (a) Linear, (b) Polynomial, (c) Radial Basis Function (RBF).

Facebook Prophet: Prophet uses a decomposable time series model [11, 22] which models three components: trend, seasonality, and holidays. They are combined additively as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_1, \quad (1)$$

where $g(t)$ is a piece-wise linear or a logistic growth curve for modelling the trend function that catches non-periodic changes in the value of the time series, $s(t)$ represents periodic changes (e.g., weekly and yearly seasonality), and $h(t)$ represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term ε_1 represents any idiosyncratic changes which are not accommodated by the model; the package assumes that ε_1 is normally distributed.

Using time as a regressor, Prophet attempts to fit several linear and nonlinear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in the Holt-Winters technique. This package frames the forecasting problem as curve-fitting rather than looking explicitly at the time based dependence of each observation within a time series. This means that it is not designed for multivariate time series.

XGBoost regression: Gradient boosting is an ensemble technique which creates a prediction model by aggregating the predictions of weak prediction models, typically decision trees. With boosting methods, weak predictors are added to the collection sequentially with each one attempting to improve upon the entire ensemble's performance.

In the XGBoost implementation [23], given a dataset with n training examples consisting of an input \mathbf{x}_i and expected output y_i , a tree ensemble model $\phi(\mathbf{x}_i)$ is defined as the sum of K regression trees $f_k(\mathbf{x}_i)$:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i).$$

To evaluate the performance of a given model, we choose a loss function $l(\hat{y}_i, y_i)$ to measure the error between the predicted value and the target value, and optionally add a regularization term $\Omega(f_k)$ to penalize overly complex trees:

$$L(\phi) = \sum_i^n l(\hat{y}_i, y_i) + \sum_k^K (\Omega(f_k)).$$

The algorithm minimizes $L(\phi)$ by iteratively introducing each f_k . Assume that the ensemble currently contains K trees. We add a new tree f_{K+1} that minimizes

$$\sum_i^n l(\hat{y}_i, y_i + f_{K+1}(\mathbf{x}_i)) + \Omega(f_k),$$

or in other words, we greedily add the tree that most improves the current model as determined by L . We train the new tree using this objective function; this is done in practice by approximating the objective function using the first and second order gradients of the loss function $l(\hat{y}_i, y_i)$ [24].

3 Implementation Methodology

We implement the four algorithms described above in Python, using the scikit-learn and Keras packages with Tensorflow as a backend [25, 26]. We used the Python implementation of Prophet [11]. Table 1 shows the configuration parameters used in the experiment for (LSTM, SVR XGboost) multivariate models and (The Facebook Prophet) univariate model. For more details regarding the implantation of the algorithms, the reader can check our longer paper [27] in that field and the packages documentations online [11, 25, 26].

Before being fed into the models, categorical features are encoded as numerical values and all features are subsequently normalized to lie in the interval [0, 1]. To test the effect of external features, we reframe the data into three different datasets for testing: one set consisting of univariate time series with no external features, one consisting of the full multivariate time series with all features, and one containing external features alone with no energy time series information. The time series

Table 1 Configuration parameters for multivariate models (LSTM, SVR XGboost) and univariate model (The Facebook Prophet)

Model	Configuration parameters
LSTM	Input layer, 50 LSTM neurons, 1 neuron output layer loss (mae), optimizer (adam), epochs (300), batch size (72)
SVR (RBF)	kernel = 'rbf', C = 1e3, gamma = 0.1
SVR (Linear)	kernel = 'linear', C = 1e3
SVR (Poly)	kernel = 'poly', C = 1e3, degree = 3
Gradient Boosting	booster(gbtree), colsample bytree (1), gamma (0) learning rate (0.1), delta step (0), max depth (3), No estimators (100)
Facebook Prophet	Default parameters, Periods (1500), freq (30T)

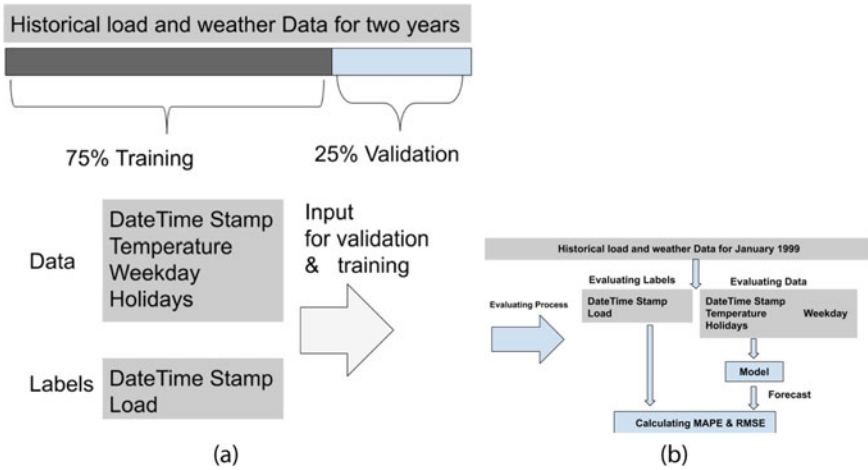


Fig. 1 (a) Training and testing processes. (b) Evaluating process

datasets are converted to input-output pairs for supervised learning by considering a sliding window 50 timesteps in which the windowed portion of the series is used to predict the next timestep.

The models are trained on each dataset with a 75/25 training/validation split as shown in Fig. 1, and evaluated on one month of reserved testing data. To avoid data leakage, we split the data in such a way that all data points in the validation set occur chronologically later than those the training set, and all data in the testing set occur after both.

The performance of the models is evaluated by two commonly used metrics in forecasting, root-mean-square error (RMSE) and mean absolute percentage error (MAPE) that are defined as the following:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \tag{2}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100. \tag{3}$$

The data used for comparison is a well-studied [4, 28] dataset obtained from the 2001 European Network on Intelligent Technologies (EUNITE) competition for electricity load forecasting [29]. This data comes from the Eastern Slovakian Electricity Corporation and spans two years from January 1, 1997 until December 31, 1998. It includes the following features: the half-hourly electricity load, the daily average temperature, and a flag signifying whether the day is a holiday. The partial autocorrelation is shown in Fig. 2; we note a one-timestep dependency as expected

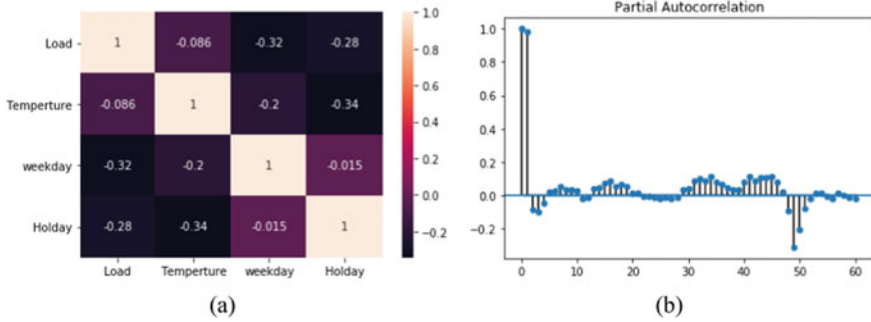


Fig. 2 (a) A correlation matrix. (b) Partial autocorrelation for the load [28]

for time series, as well as a spike around 48 timesteps corresponding to daily cycles. We also notice a spike at 336 timesteps corresponding to weekly cycles. For this reason the important past values for our time series model to incorporate are a lag of 1, and ideally 48 and 336 as well.

A correlation coefficient of -0.8676 between the daily peak load and the daily average temperature indicates a strong relationship between the electrical load and weather conditions [28]. Analysis of the dataset shows that the load generally reduces on holidays and weekends [4, 28], likely due to businesses shutting down. This varies depending on the specific holiday; on Christmas or New Year, for example, electricity consumption is affected more than on other holidays. Based on these observations, we choose to use as input features for our experiments the past loads, daily temperature, the time of day, month, day of the week and whether the day is a holiday [4, 28]. These features are encoded as numerical or binary values and normalized to lie in the range $[0, 1]$ using the `MinMaxScaler` from `scikit-learn`, while categorical features are one-hot encoded using `LabelEncoder` from `scikit-learn`.

4 Experimental Results

A one-month forecast obtained from the four models is shown in Fig. 3 for 100 timesteps (50h). Qualitatively, it can be seen from the figure that the forecast is fairly accurate for the LSTM, SVR (RBF, linear), and XGboost models, while being considerably worse for SVR (polynomial). The figure also shows that the forecasts obtained from Prophet consistently overestimate the actual value, while at the same time not capturing small-scale variations in the load behaviour. We believe that the superior performance of the LSTM, SVR, and XGboost models is due to the incorporation of multivariate data. Note that despite the external features provided to the models are measured daily (temperature is provided as a daily average), the multivariate models still exhibit superior performance.

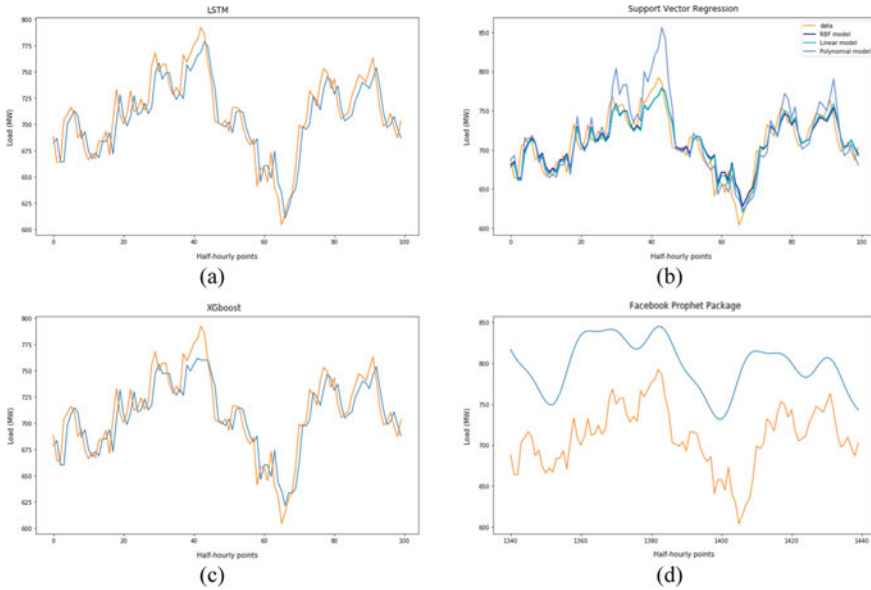


Fig. 3 Predictions (blue) compared with actual values (orange) for 100 time steps in January 1999 using (a) LSTM, (b) SVR, (c) XGboost, (d) Facebook Prophet Package

Table 2 MAPE and RMSE for multivariate models (LSTM, SVR XGboost) and (The Facebook Prophet) univariate model

	LSTM	SVR RBF Ker.	SVR Linear Ker.	SVR Poly Ker.	Gradient Boosting	Facebook Prophet
MAPE	1.51	2.1	2.1	4.0	2.1	14.4
RMSE	13.5	17.6	17.9	36.7	16.64	102.2

Table 2 shows the RMSE and MAPE values for each model while predicting the half-hourly load for one month. The results for multivariate models are as following for the LSTM model which obtains the highest accuracy with a MAPE value of 1.51% and RMSE value of 13.5 MW, followed by SVR (RBF, Linear) and XGBoost with MAPE values of 2.1% and RMSE values of 17.6 MW, 17.9 MW, and 18.2 MW respectively. While the lowest accuracy is for the Facebook Prophet univariate model with a MAPE value of 14.4% and RMSE value of 102.2 MW

To estimate the contribution of the external features on multivariate models accuracy, we conducted the same experiment for the LSTM, SVR, and XGboost models without using past power consumption as an input feature. Table 3 shows the RMSE and MAPE values for each model while predicting the half-hourly load for one month. The LSTM model obtains the highest accuracy with a MAPE value of 6.1% and RMSE value of 51.161 MW, followed by SVR (Poly) and XGBoost with MAPE values of 6.4 and 7.5%, respectively. We note that multivariate models still achieve

Table 3 MAPE and RMSE for multivariate models (LSTM, SVR XGBoost) and univariate model (Prophet) without using past power consumption as input feature

	LSTM	SVR RBF Ker.	SVR Linear Ker.	SVR Poly Ker.	XGBoost	Facebook Prophet
MAPE	6.1	16.0	12.1	6.4	7.5	14.4
RMSE	51.161	128.355	96.036	52.863	63.135	102.2

reasonable accuracy and outperform the univariate model even without using past power consumption as an input feature.

5 Conclusion

External features, even when provided on longer timescales than the time series of interest, can be used to improve prediction accuracy. In this work, we compare four forecasting algorithms for time series—LSTM, SVR, XGBoost, and the Prophet package—for the problem of short-term energy load forecasting. We show that despite the external features of interest (e.g., temperature and holidays) being measured on a daily basis, they considerably increase the accuracy of the forecast for multivariate models as compared to the univariate model. Even when past values are not provided to the model, the models achieve reasonable accuracy based only on these external features and the time of day.

As future work, we intend to use datasets from other areas such as finance and medical applications to investigate the consistency of algorithm performance. We will also consider to develop new computational models that would take the advantages of both multivariate and univariate time series algorithms.

Acknowledgements The project was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Amasyali, K., El-Gohary, N.M.: A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **81**, 1192–1205 (2018)
2. Bianchi, F.M., Maiorino, E., Kampffmeyer, M.C., Rizzi, A., Jenssen, R.: An overview and comparative analysis of recurrent neural networks for short term load forecasting. [arXiv:1705.04378](https://arxiv.org/abs/1705.04378) (2017)
3. Saleh, A.I., Rabie, A.H., Abo-Al-Ez, K.M.: A data mining based load forecasting strategy for smart electrical grids. *Adv. Eng. Inform.* **30**(3), 422–448 (2016)
4. Chen, B.-J., Chang, M.-W., Lin, C.-J.: Load forecasting using support vector machines: a study on EUNITE competition 2001. *IEEE Trans. Power Syst.* **19**(4), 1821–1830 (2004)

5. Dannecker, L.: *Energy Time Series Forecasting: Efficient and Accurate Forecasting of Evolving Time Series from the Energy Domain*. Springer (2015)
6. Jiang, F., Yang, X., Li, S.: Comparison of forecasting India's energy demand using an MGM, ARIMA model, MGM-ARIMA model, and BP neural network model. *Sustainability* **10**(7), 2225 (2018)
7. Yuan, C., Liu, S., Fang, Z.: Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM (1, 1) model. *Energy* **100**, 384–390 (2016)
8. Narayan, A., Hipel, K.W.: Long short term memory networks for short-term electric load forecasting. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1050–1059, Banff Center, Banff, Canada, 5–8 Oct 2017
9. Chen, Y., Xu, P., Chu, Y., Li, W., Wu, Y., Ni, L., Bao, Y., Wang, K.: Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings. *Appl. Energy* **195**, 659–670 (2017)
10. Li, G.Y., Li, W., Tian, X.L., Che, Y.F.: Short-term electricity load forecasting based on the XGBoost algorithm. *Smart Grid* **07**, 274–285 (2017)
11. Taylor, S.J., Letham, B.: Forecasting at scale. *Am. Stat.* **72**(1), 37–45 (2018)
12. Montgomery, D.C., Jennings, C.L., Kulahci, M.: *Introduction to Time Series Analysis and Forecasting*. Wiley Series in Probability and Statistics. Wiley (2015)
13. Chatfield, C.: *Time-Series Forecasting*. CRC Press (2000)
14. Tsay, R.S.: *Multivariate Time Series Analysis: With R and Financial Applications*. Wiley (2013)
15. Kanchymalay, K., Salim, N., Sukprasert, A., Krishnan, R., Hashim, U.R.: Multivariate time series forecasting of crude palm oil price using machine learning techniques. In: *IOP Conference Series: Materials Science and Engineering*, vol. 226, p. 012117. IOP Publishing (2017)
16. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**(5–6), 602–610 (2005)
17. Olah, C.: Understanding LSTM networks. GITHUB blog, posted on 27 Aug 2015
18. Gamboa, J.C.B.: Deep learning for time-series analysis. [arXiv:1701.01887](https://arxiv.org/abs/1701.01887) (2017)
19. Zhu, L., Laptev, N.: Deep and confident prediction for time series at Uber. [arXiv:1709.01907](https://arxiv.org/abs/1709.01907) (2017)
20. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004)
21. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer Science & Business Media (2013)
22. Harvey, A.C., Peters, S.: Estimation procedures for structural time series models. *J. Forecast.* **9**(2), 89–108 (1990)
23. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM (2016)
24. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **28**(2), 337–407 (2000)
25. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. [arXiv:1603.04467](https://arxiv.org/abs/1603.04467) (2016)
26. Chollet, F., et al.: Keras. <https://keras.io> (2015)
27. Selim, M., Quinsey, P., Feng, W., Zouh, R.: Uncertainty for energy forecasting using Bayesian deep learning. Submitted to the *J. Math. Found. Comput. (MFC)* (2020)
28. Nagi, J., Yap, K.S., Nagi, F., Tiong, S.K., Ahmed, S.K.: A computational intelligence scheme for the prediction of the daily peak load. *Appl. Soft Comput.* **11**(8), 4773–4788 (2011)
29. EUNITE: Eunite electricity load forecast 2001 competition. In: *Proceedings of EUNITE*, Dec 2001

Toral Diffeomorphisms Induce Quantum Superoperators via TAQS



Artur Sowa

Abstract We propose a new method for adapting (perturbing) models of quantum observables. The method is dubbed TAQS as it is based on *toral automorphisms* (diffeomorphisms) and the *Q-transform*, which together induce *superoperators* acting on observables. We demonstrate via examples that TAQS perturbations often lead to radical changes in the observables' structure and spectra. This is a preliminary exploration in which emphasis is put on connections with some exciting canonical topics (the almost Mathieu operators), and with recent trends in the study of quantum metamaterials (fractal-structured operators).

Keywords Quantum theory · Superoperators, Q-transform · Automorphisms of a torus

1 Introduction

A study of perturbations of Hamiltonians plays an important role in Quantum Theory and in its applications to Condensed Matter Physics, Materials Science, Chemistry, Synchrotron Science, etc. The traditional approach goes back to the 1930s work of E. Wigner, H. A. Jahn, E. Teller, and others, and focuses on the action of symmetry groups, and the concept of symmetry breaking. We propose an *alternative approach* dubbed TAQS. The TAQS method is based on an observation that a Hilbert space operator can be perturbed via topological automorphisms of the two-torus. In other words, diffeomorphisms of the torus are interpreted as superoperators acting in spaces of observables. Here, we will examine some effects related to special diffeomorphisms: the cyclic shifts and automorphisms identified with the elements of $GL(2, \mathbb{Z})$.

The proposed construction utilizes the Q-transform, [9], which identifies the generalized functions on the torus with Hilbert space operators. Since diffeomorphisms

A. Sowa (✉)

Department of Mathematics and Statistics, Center for Quantum Topology and Its Applications (quanTA), University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK S7N5E6, Canada
e-mail: sowa@math.usask.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_41

441

of the torus act on functions (distributions), they automatically act on operators. An operator obtained by perturbing a self-adjoint operator is also self-adjoint. However, it may have dramatically different properties, e.g. perturbations of the identity operator by some toral rotations have nontrivial spectra. Even more strikingly, a perturbation of the simple discrete second derivative is unitarily equivalent to the almost Mathieu operator, see Example 3 in Sect. 3. The latter plays a role in the modelling of disordered electronic transport, and has been continually researched by physicists and mathematicians for decades, [3, 7]. In particular, it was conjectured that (under some conditions) the spectrum of this operator is a Cantor set, [8]. This conjecture, known as the Ten Martini Problem, was settled positively in 2009, [2].

TAQS may also be used in investigations of the structure of materials. The envisioned method involves switching from the bottom-up to a top-down approach. Namely, in most of the classical science one constructs models starting from the fundamental principles and gradually adding structure, see e.g. [5]. In contrast, modern computational optimization methods enable one to construct models based on the best fit to (whatever type of) data, notwithstanding that the results need not be easy to interpret via the fundamental principles. A combination of these concepts suggests a method for finding the best model among the specific adaptations. The starting point is construction of special superoperators put forth here, which are based on toral shifts and automorphisms obtained via $GL(2, \mathbb{Z})$. Since the shifts are in numerical practice discrete, and the group of lattice automorphisms is finitely generated, this framework is amenable to AI explorations. We hope to provide more details in forthcoming publications.

2 The Method of TAQS

Any diffeomorphism of the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$ acts on functions or distributions via the change of variable. On the other hand, one can identify functions (and distributions) with Hilbert space operators via the Q-transform, [9]. In this way, one obtains a representation of the group of diffeomorphisms in the space of operators. In what follows we discuss this construction in more detail.

2.1 Q-Transform Mediated Equivalence Between Functions and Operators

For an absolutely integrable function $f : \mathbb{T} \rightarrow \mathbb{R}$, its Fourier coefficients $[\hat{f}(k, l)]_{(k, l) \in \mathbb{Z}^2}$ are defined via:

$$\hat{f}(k, l) = \iint f(x, y) e^{-2\pi i(kx + ly)} dx dy.$$

The definition admits a well-known extension for distributions. The set of all real-valued distributions on \mathbb{T} will be denoted \mathcal{S} . Next, we fix a Hilbert space \mathbb{H} with a distinguished unitary basis,¹ say, $(e_k)_{k \in \mathbb{Z}}$. We will use these data to construct a quantum observable, say, $\mathcal{H} : \mathbb{H} \rightarrow \mathbb{H}$. We do so by prescribing the observable's matrix elements, i.e. $h_{k,l} = \mathcal{H}(k, l) = \langle e_k | \mathcal{H} e_l \rangle$. Specifically, we set

$$h_{k,l} = \begin{cases} \hat{f}(k, l) & \underline{\text{if}} & k < l \\ \hat{f}(l, k)^* & \underline{\text{if}} & k > l \\ \sqrt{2} \Im \hat{f}(k, k) & \underline{\text{if}} & l = k < 0 \\ \hat{f}(0, 0) & \underline{\text{if}} & l = k = 0 \\ \sqrt{2} \Re \hat{f}(k, k) & \underline{\text{if}} & l = k > 0 \end{cases} \quad (1)$$

Clearly, the matrix \mathcal{H} is self-adjoint, i.e. $h_{k,l} = h_{l,k}^*$. This operation is invertible as indeed:

$$\hat{f}(k, l) = \begin{cases} h_{k,l} & \underline{\text{if}} & k < l \\ h_{-k,-l}^* & \underline{\text{if}} & k > l \\ \frac{1}{\sqrt{2}} (h_{-k,-k} + i h_{k,k}) & \underline{\text{if}} & l = k < 0 \\ h_{0,0} & \underline{\text{if}} & l = k = 0 \\ \frac{1}{\sqrt{2}} (h_{k,k} - i h_{-k,-k}) & \underline{\text{if}} & l = k > 0 \end{cases} \quad (2)$$

We write

$$\mathcal{H} = Q[f], \quad f = Q^{-1}[\mathcal{H}] \quad \text{and} \quad \mathcal{H} = S[\hat{f}], \quad \hat{f} = S^{-1}[\mathcal{H}].$$

Operation Q , referred to as the Q-transform, is a composition of the Fourier transform with the symmetry change S . It is linear and invertible. We will refer to $\mathcal{O} = Q\mathcal{S}$ as the space of observables. In other words, \mathcal{O} is the linear space of all operators in \mathbb{H} (with the distinguished basis) that Q^{-1} maps into real-valued distributions. Some subspaces of \mathcal{O} have special properties. In particular, the following useful observation follows directly from (1) and the Parseval's theorem:

Fact 1 *The Q-transform is a unitary map from the space of real square-integrable functions onto the space of self-adjoint Hilbert-Schmidt operators.*

¹ It is important that the basis index set consist of integers (rather than natural numbers). The finite-dimensional version of the Q-transform is also easy to interpret but requires an odd number of indices, [9].

In fact, more is true: the Q-transform enables the definition of Sobolev classes of observables (relative to the choice of basis in \mathbb{H}). This concept can be applied in the analysis of quantum dynamics, [9], but that is not our focus.

2.2 Toral Automorphisms Acting on Functions and Operators

Any diffeomorphism of the torus, say, $\Phi : \mathbb{T} \rightarrow \mathbb{T}$, acts on functions via the change of variable $f \mapsto f \circ \Phi$. The action is naturally extended to \mathcal{S} (as pullback on distributions). The Q-transform, in turn, extends it to action on \mathcal{O} . In other words, Φ gives rise to a superoperator acting on operators, $Q[f] \mapsto Q[f \circ \Phi]$, denoted Σ_Φ . We will refer to these superoperators and their compositions as TAQS. Clearly, $f \mapsto f \circ \Phi$ is unitary in the real Hilbert space $L_2(\mathbb{T})$, whenever Φ is measure preserving. Fact 1 implies:

Fact 2 *When Φ is a measure preserving diffeomorphism, the superoperator Σ_Φ is a unitary map in the space of self-adjoint Hilbert-Schmidt operators.*

One can interpret the space of all Hilbert-Schmidt operators as the space of quantum states. The action of superoperators Σ_Φ induced by any measure preserving diffeomorphisms is easily extended to this space and remains unitary. Thus, this structure furnishes an alternative model of the standard quantum mechanics. It additionally brings on board a constellation of topological as well as chaos-theoretic questions. Such questions can also be addressed directly via the space of complex square integrable functions on the torus. However, the setting enabled by the Q-transform brings an essentially new element into the scope: the effect of superoperators on the structure and, in particular, on the spectra of observables. We will examine it via several calculable examples that stem from special diffeomorphisms:

- Cyclic shifts, i.e. for $[\alpha, \beta] \in \mathbb{T}$, one has the map $(x, y) \mapsto T(x, y) = T_{[\alpha, \beta]}(x, y) = (x + \alpha, y + \beta) \bmod 1$.
- Diffeomorphisms induced by matrices $M \in GL(2, \mathbb{Z})$, i.e. for

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad a, b, c, d \in \mathbb{Z}, \quad \text{and } ad - bc = \pm 1, \tag{3}$$

one has the map $(x, y) \mapsto \Phi_M(x, y) = (ax + by, cx + dy) \bmod 1$.

We will refer to the group of automorphisms generated by \mathbb{T} and $GL(2, \mathbb{Z})$ as $\text{Aut}_{\mathbb{T}}$. It is convenient to introduce the following notation:

$$f_M = f \circ \Phi_M^{-1}, \quad f_T = f \circ T^{-1}. \tag{4}$$

The corresponding superoperators, denoted for brevity Σ_M, Σ_T , are then expressed via

$$\Sigma_M[\mathcal{H}] = \mathcal{H}_M := Q[Q^{-1}[\mathcal{H}]_M], \quad \Sigma_T[\mathcal{H}] = \mathcal{H}_T := Q[Q^{-1}[\mathcal{H}]_T]. \quad (5)$$

This defines a representation of $\text{Aut}_{\mathbb{T}}$ in the vector space \mathcal{O} .

2.3 Explicit Formulas for the Special TAQS

We present a straightforward result, Proposition 1, that makes explicit some basic features of TAQS. It is convenient to use the following notation:

$$U_+ = \{(k, l) \in \mathbb{Z}^{\times 2} : k < l\}, \quad U_- = \{(k, l) \in \mathbb{Z}^{\times 2} : k > l\}, \\ D_+ = \{(k, l) \in \mathbb{Z}^{\times 2} : k = l > 0\}, \quad D_- = \{(k, l) \in \mathbb{Z}^{\times 2} : k = l < 0\}.$$

Thus, the lattice \mathbb{Z}^2 is partitioned into five disjoint sets: U^+, U_-, D_+, D_- , and $\{(0, 0)\}$.

Proposition 1 *The matrix coefficients of $\mathcal{H}_T, \mathcal{H}_M$ are obtained from those of \mathcal{H} as follows:*

1. $\mathcal{H}_T(0, 0) = \mathcal{H}(0, 0) = \mathcal{H}_M(0, 0)$.
2. For $T = [\alpha, \beta]$, we have:

$$\mathcal{H}_T(k, l) = \begin{cases} h_{k,l} \exp[-2\pi i(k\alpha + l\beta)] & (k, l) \in U_+ \\ h_{k,l} \exp[2\pi i(k\beta + l\alpha)] & (k, l) \in U_- \end{cases} \quad (6)$$

Thus, the off-diagonal entries of \mathcal{H}_T differ from those of \mathcal{H} only in phase. Furthermore, for $k > 0$:

$$\begin{bmatrix} \mathcal{H}_T(k, k) \\ \mathcal{H}_T(-k, -k) \end{bmatrix} = \begin{bmatrix} \cos 2\pi k(\alpha + \beta) & -\sin 2\pi k(\alpha + \beta) \\ \sin 2\pi k(\alpha + \beta) & \cos 2\pi k(\alpha + \beta) \end{bmatrix} \begin{bmatrix} h_{k,k} \\ h_{-k,-k} \end{bmatrix} \quad (7)$$

3. Let M be as in (3), and let $k_1 = ak_0 + cl_0, l_1 = bk_0 + dl_0$. If $(k_0, l_0) \in U_+$, we have

$$\det M \cdot \mathcal{H}_M(k_0, l_0) = \begin{cases} h_{k_1, l_1} & (k_1, l_1) \in U_+ \\ h_{-k_1, -l_1}^* & (k_1, l_1) \in U_- \\ \frac{1}{\sqrt{2}}(h_{k_1, k_1} - ih_{-k_1, -k_1}) & (k_1, l_1) \in D_+ \\ \frac{1}{\sqrt{2}}(h_{-k_1, -k_1} + ih_{k_1, k_1}) & (k_1, l_1) \in D_- \end{cases} \quad (8)$$

The coefficients $\mathcal{H}_M(k_0, l_0)$ for $(k_0, l_0) \in U_-$ are also obtained from (8) via self-adjointness. Furthermore, if $(k_0, l_0) \in D_+$, we have

$$\det M \cdot \mathcal{H}_M(k_0, l_0) = \begin{cases} \sqrt{2} \Re h_{k_1, l_1} & (k_1, l_1) \in U_+ \\ \sqrt{2} \Re h_{-k_1, -l_1}^* & (k_1, l_1) \in U_- \\ h_{k_1, k_1} & (k_1, l_1) \in D_+ \\ h_{-k_1, -k_1} & (k_1, l_1) \in D_- \end{cases} \quad (9)$$

Finally, if $(k_0, l_0) \in D_-$, we have

$$\det M \cdot \mathcal{H}_M(k_0, l_0) = \begin{cases} \sqrt{2} \Im h_{k_1, l_1} & (k_1, l_1) \in U_+ \\ \sqrt{2} \Im h_{-k_1, -l_1}^* & (k_1, l_1) \in U_- \\ -h_{-k_1, -k_1} & (k_1, l_1) \in D_+ \\ h_{k_1, k_1} & (k_1, l_1) \in D_- \end{cases} \quad (10)$$

Proof First, the invariance of the coefficient at $(0, 0)$ follows directly from the definitions.

Second, let $f = Q^{-1}[\mathcal{H}]$. (6–7) follow from (1–2) via the straightforward formula $\hat{f}_T(k, l) = \hat{f}(k, l) \exp[-2\pi i(k\alpha + l\beta)]$.

Finally, formulas (8–10) are obtained in the same way via identity $\hat{f}_M(k, l) = (\det M) \hat{f}(ak + cl, bk + dl)$. \square

Nontrivial phases in the off-diagonal terms of Hamiltonians (such as in (6)) turn up in models of electrons hopping on a lattice with transversal magnetic field.² This is of relevance to phenomena such as the Quantum Hall Effect, e.g. [4, 11]. At the same time, an application of Σ_M could be interpreted as a redesign of the hopping constraints (e.g. nearest neighbour vs. mid-range).

3 A few Examples of the Effect of TAQS

It is easily seen from (6–7), that the action by shift $[-\alpha, \alpha]$ is a unitary equivalence; for more details see [9]. However, that is not the case for other choices of the shift. Moreover, it is easily seen that for a rank one operator, say, $\mathcal{H} = |\psi\rangle\langle\psi|$, the operators $\Sigma_T[\mathcal{H}]$ or $\Sigma_M[\mathcal{H}]$ are generally not of rank one.

² More precisely, the direct sum of a discrete family of Hamiltonians is needed to model electron hopping on \mathbb{Z}^2 lattice.

A self-adjoint Hilbert-Schmidt operator is compact and so its spectrum is either a finite set or a countably infinite set with an accumulation point at 0. The TAQS can modify such a spectrum, e.g. change the value of the eigenvalues or even their number. We examine two explicit examples as that using finite rank operators (examples 1, 2). We also demonstrate that when applied to a non Hilbert-Schmidt operator, a TAQS superoperator can change the spectrum’s topology in a highly nontrivial way (Example 3). We also discuss a substantial change of an operator’s structure under a specific TAQS action (Example 4). Finally, we display explicit examples of eigenvectors of TAQS (Example 5). In all examples the Hilbert space \mathbb{H} is infinite-dimensional (except for the numerical illustration in Example 4).

Example 1 *Rank and eigenvalue modulation via Σ_T .* Let $\tilde{\sigma}_\alpha$ with $\alpha = 1, 2, 3$ be an operator defined by the action of the Pauli matrix σ_α in $\text{span}\{e_0, e_1\} \subset \mathbb{H}$, whereas $\tilde{\sigma}_\alpha e_n = 0$ for all $n \neq 0, 1$. We select the zero-trace Hamiltonian, $\mathcal{H} = p_1\tilde{\sigma}_1 + p_2\tilde{\sigma}_2 + p_3\tilde{\sigma}_3$, i.e.

$$\mathcal{H} = \begin{pmatrix} p_3 & p_1 - ip_2 \\ p_1 + ip_2 & -p_3 \end{pmatrix} \text{ in } \text{span}\{e_0, e_1\} \tag{11}$$

Normalizing $p_1^2 + p_2^2 + p_3^2 = 1$, we find that the eigenvalues of \mathcal{H} are ± 1 (in addition to 0), regardless of the parameters. It follows from (6–7) that $\mathcal{H}_T = \Sigma_T[\mathcal{H}]$ is nonzero only in the invariant subspace $\text{span}\{e_{-1}, e_0, e_1\}$, wherein:

$$\mathcal{H}_T = \begin{pmatrix} -p_3 \sin 2\pi\theta & 0 & 0 \\ 0 & p_3 & (p_1 - ip_2)e^{-2\pi i\beta} \\ 0 & (p_1 + ip_2)e^{2\pi i\beta} & -p_3 \cos 2\pi\theta \end{pmatrix}. \tag{12}$$

Here, $\theta = \alpha + \beta$. The eigenvalues of this block matrix are readily found explicitly; they are independent of β and periodic in θ . There are generically three distinct eigenvalues but also some level-crossings. If $p_3 = 0$ or $\theta = 0$, then \mathcal{H}_T is unitarily equivalent to \mathcal{H} .

Example 2 *Rank augmentation via Σ_M .* We now consider a perturbation via M given in (14) of the Hamiltonian (11). It follows from (8–9–10), that \mathcal{H}_M is nonzero only in the invariant subspace $\text{span}\{e_0, e_1, e_2\}$, wherein:

$$\mathcal{H}_M = \begin{pmatrix} p_3 & p_1 - ip_2 & 0 \\ p_1 + ip_2 & 0 & -p_3/\sqrt{2} \\ 0 & -p_3/\sqrt{2} & 0 \end{pmatrix}. \tag{13}$$

The characteristic polynomial depends only on p_3 (due to the normalization condition). It is easily seen that for every $p_3 \in [-1, 1]$, the matrix has three distinct real eigenvalues. Again, the rank of the operator does not increase if $p_3 = 0$.

Example 3 *The shifted identity and the almost Mathieu operator.* Let $I : \mathbb{H} \rightarrow \mathbb{H}$ be the identity operator³. Let T be a shift of the torus by $[\alpha, \beta]$, and denote $\theta = \alpha + \beta$. It follows from (7) that the operator I_T is diagonal, and

$$I_T(k, k) = \sqrt{2} \cos(2\pi k\theta + \pi/4).$$

Thus, when θ is irrational, the spectrum of I_T is the interval $[-\sqrt{2}, \sqrt{2}]$ and, when it is rational, it is a finite subset of this interval. Furthermore, let $R : \mathbb{H} \rightarrow \mathbb{H}$ be the right-shift operator, defined via $R[e_n] = e_{n+1}$. Then, with some real parameter μ ,

$$\mathcal{H}_{\mu,\theta} = \sqrt{2} \mu I_T + R + R^\dagger$$

is the famous *almost Mathieu operator*.

One verifies directly that acting with Σ_T on the tridiagonal matrix $\Delta_\mu = \mu I + R + R^\dagger$ introduces variable phase in the off-diagonal coefficients. However, taking $\alpha = \beta$, we find that the operator $\Sigma_T[\Delta_\mu]$ is unitarily equivalent with the almost Mathieu operator, namely:

$$\Sigma_T[\Delta_\mu] = U \mathcal{H}_{\mu,2\alpha} U^\dagger,$$

where U is a diagonal unitary matrix with $U(k, k) = \exp(2\pi i k^2 \alpha)$. Thus, the TAQS action transforms Δ_μ whose spectrum is $[-2 + \mu, 2 + \mu]$ into $\Sigma_T[\Delta_\mu]$ whose spectrum is topologically the Cantor set (for $\mu \neq 0$ and irrational α), [2].

Example 4 *The effect of $GL(2, \mathbb{Z})$ on the structure of couplings.* Consider a matrix \mathcal{H} that has nonzero coefficients $h_{k,l}$ aligning along lines $k - l = \gamma$ for some values of the constant γ , as in the example given in Fig. 1a. Next, let us choose the map

$$M = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}. \tag{14}$$

Applying formula (8) we find that \mathcal{H}_M has nonzero coefficients aligning along the lines $2k - l = \gamma'$ for some values of γ' , as seen in Fig. 1b. Examples as \mathcal{H} used here have recently been introduced for the purpose of analyzing quantum metamaterials, [10]. In fact, the concentration of coefficients along lines reflects the structure of certain physical couplings. The TAQS action via M redesigns those couplings.

The numerical result displayed here is based on the discrete version of the Q-transform, [9], and a discrete version of an automorphism of the torus. Specifically, for a function $f : \mathbb{T} \rightarrow \mathbb{R}$ represented by an $N \times N$ matrix, the matrix of f_M is computed via recalculating indices:

³ It is easily seen that $I \in \mathcal{O}$, as is the case for other operators considered in this section.

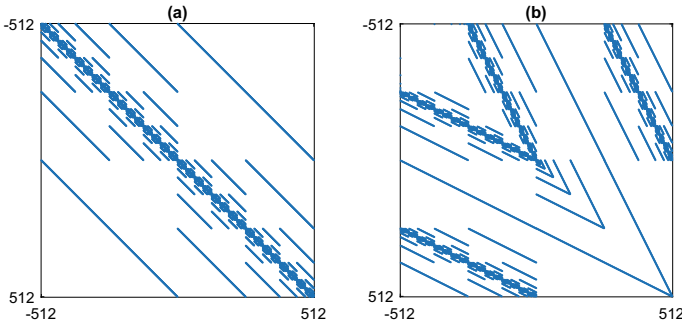


Fig. 1 **a** The support (i.e. location of the nonzero entries) of the original matrix $\mathcal{H} = 0 \oplus \sum_k \frac{1}{2^k} \sigma_x^{(k)}$, where the sum is finite. Note that the nonzero entries align along certain lines with slope 1. **b** The support of \mathcal{H}_M where M is as specified in (14). The nonzero entries align along certain lines with the slope 2 and 1/2. From the physical point of view this represents a rearrangement of couplings between qubits. *Note: This example is obtained numerically via a discrete model of the torus. The choice of matrix indexing (by integers centered at 0) is natural for an application of the discrete Q-transform, but needs to be recalibrated (to positive integers) when computing the toral automorphism. The discrete results are different in some details than the infinite-dimensional version discussed in Proposition 1*

```

ind = mod(M^(-1) * [1:N; 1:N], N);
ind = ind + N*(ind == 0); % replace 0 by N
f_M = f(ind(1, :), ind(2, :));
    
```

Nevertheless, the highlighted features are similar as in Proposition 1.

Example 5 *Nontrivial invariance.* We give an example of a nontrivial operator that remains invariant under a nontrivial automorphism. Consider a sequence of real numbers $(x_n)_{n \in \mathbb{N}}$ with $x_1 = 0$. Define \mathcal{H} as follows: For $0 < k, l$ we set

$$\mathcal{H}(k, l) = \begin{cases} -i x_{k/l} & \text{if } l|k \\ i x_{l/k} & \text{if } k|l \end{cases} \tag{15}$$

All other coefficients are set to zero. The coefficient $i x_n$ occurs repeatedly along the line $-nk + l = 0$. At the same time, $-i x_n$ occurs repeatedly along the line $k - nl = 0$. (Note that with some assumptions on the growth of $|x_n|$, e.g. if the sequence is bounded, we have $\mathcal{H} \in \mathcal{O}$.) Let us consider Σ_{σ_1} induced by the first Pauli matrix $\sigma_1 \in GL(2, \mathbb{Z})$. The corresponding TAQS action intertwines the two sets of lines, and (8) yields

$$\mathcal{H}_{\sigma_1} = \Sigma_{\sigma_1}[\mathcal{H}] = \mathcal{H}.$$

It is known, [1], that Φ_{M_1} and Φ_{M_2} are topologically conjugate if and only if M_1 and M_2 are similar matrices, i.e. there exists⁴ $M \in GL(2, \mathbb{Z})$, such that $MM_1 = M_2M$.

⁴ Also, the problem of similarity of matrices over $GL(2, \mathbb{Z})$ is nontrivial but well understood, [6].

That implies similarity of the corresponding superoperators $\Sigma_M \Sigma_{M_1} = \Sigma_{M_2} \Sigma_M$. In particular, if \mathcal{H} is invariant under Σ_{M_1} , then \mathcal{H}_M is an eigenvector of Σ_{M_2} . In particular, taking \mathcal{H} as in (15), and arbitrary $M \in GL(2, \mathbb{Z})$, we observe that

$$\Sigma_{M_2}[\mathcal{H}_M] = \mathcal{H}_M \quad \text{wherein } M_2 = M\sigma_1 M^{-1}.$$

These are the first examples of eigenvectors of TAQS.

Acknowledgements My thinking about the matters presented here has been influenced by the following people: John-Carl Bermodes, Robert Green, Natalia Janson, Bing-Zhao Li, Robert Moody, Raymond Spiteri, and Alexandre Zagoskin.

References

1. Adler, R.L., Weiss, B.: Similarity of automorphisms of the torus. *Mem. Amer. Math. Soc.* **98** (1970)
2. Avila, A., Jitomirskaya, S.: The ten martini problem. *Ann. Math.* **170**, 303–342 (2009)
3. Jitomirskaya, S., Liu, W.: Universal hierarchical structure of quasiperiodic eigenfunctions. *Ann. Math.* **187**, 721–776 (2018)
4. Fradkin, E., Kohmoto, M.: Quantum Hall effect and geometrical localization of electrons on lattices. *Phys. Rev. B* **35**, 6017–6023 (1987)
5. Green, R.J., Haverkort, M.W., Sawatzky, G.A.: Bond disproportionation and dynamical charge fluctuations in the perovskite rare-earth nickelates. *Phys. Rev. B* **94**, 195127 (2016)
6. Henninger, J.P.: Factorization and Similarity in $GL(2, \mathbb{Z})$. *Lin. Alg. App.* **251**, 223–237 (1997)
7. Last, Y.: Spectral theory of Sturm-Liouville operators on infinite intervals: a review of recent developments. In: Amrein, W.O., Hinz, A.M., Pearson, D.B. (eds.) *Sturm-Liouville Theory*, pp. 99–120. Past and Present, Birkhuser, Basel (2005). MR 2145079 Zbl 1098.39011
8. Simon, B.: Schrödinger operators in the twenty-first century. In: Fokas, A., Grigorian, A., Kibble, T., Zegarlinski, B. (eds.) *Mathematical Physics*, pp. 283–288. Imperial College Press, London (2000)
9. Sowa, A.: A nonlocal transform to map and track quantum dynamics. *J. Phys. A Math. Theor.* **52**, 305301 (2019)
10. Sowa, A., Zagoskin, A.: An exactly solvable quantum-metamaterial type model. *J. Phys. A Math. Theor.* **52**, 395304 (2019)
11. Thouless, D.J., Kohmoto, M., Nightingale, M.P., den Nijs, M.: Quantized Hall conductance in a two-dimensional periodic potential. *Phys. Rev. Lett.* **49**, 405–408 (1982)

Optimal Time Decay Rates for a Chemotaxis Model with Logarithmic Sensitivity



Yanni Zeng and Kun Zhao

Abstract We consider a Keller-Segel type chemotaxis model with logarithmic sensitivity and density-dependent production/consumption rate. It is a 2×2 reaction-diffusion system describing the interaction of cells and a chemical signal. We study Cauchy problem for the original system and its transformed system, which is one of hyperbolic-parabolic conservation laws. In both cases of diffusive and non-diffusive chemical, we obtain optimal L^2 time decay rates for the solution. Our results improve those in Li et al. (Nonlinearity 28:2181-2210, 2015 [5]), Martinez et al. (Indiana Univ Math J 67:1383-1424, 2018 [7]).

Keywords Conservation laws · Hyperbolic-parabolic · Reaction-diffusion · Asymptotic behavior · Time decay

1 Introduction

In this paper we consider Cauchy problem of a Keller-Segel type chemotaxis model:

$$\begin{cases} s_t = \varepsilon s_{xx} - \mu us - \sigma s, \\ u_t = Du_{xx} - \chi[u(\ln s)_x]_x, \end{cases} \quad x \in \mathbb{R}, \quad t > 0, \quad (1)$$
$$(s, u)(x, 0) = (s_0, u_0)(x), \quad x \in \mathbb{R}. \quad (2)$$

Here the unknown functions $s = s(x, t)$ and $u = u(x, t)$ are the concentration of a chemical signal and the density of a cellular population, respectively. The constant system parameters are $\varepsilon \geq 0$, $\mu \neq 0$, $\sigma \geq 0$, $D > 0$ and $\chi \neq 0$, standing for

Y. Zeng (✉)

Department of Mathematics, University of Alabama at Birmingham, Birmingham, USA

e-mail: ynzeng@uab.edu

K. Zhao

Department of Mathematics, Tulane University, New Orleans, USA

e-mail: kzhao@tulane.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_42

451

the diffusion coefficient of chemical signal, coefficient of density-dependent production/consumption rate of chemical signal, natural degradation rate of chemical signal, diffusion coefficient of cellular population, and coefficient of chemotactic sensitivity, respectively. Equation (1) describes the movement of a cellular population in response to a chemical signal, while both entities are naturally diffusing and producing/degrading in the local environment.

Equation (1) is a system of reaction-diffusion equations. It can be transformed into a system of hyperbolic-parabolic conservation laws by the inverse Hopf-Cole transformation [4]:

$$v = (\ln s)_x = \frac{s_x}{s}. \tag{3}$$

The new system under the variables v and u reads:

$$\begin{cases} v_t + (\mu u - \varepsilon v^2)_x = \varepsilon v_{xx}, \\ u_t + \chi (uv)_x = Du_{xx}. \end{cases} \tag{4}$$

Throughout this paper we assume

$$\chi\mu > 0, \tag{5}$$

which includes two scenarios: $\chi > 0$ and $\mu > 0$, or $\chi < 0$ and $\mu < 0$. The former is interpreted as cells are attracted to and consume the chemical. The latter describes cells depositing the chemical to modify the local environment for succeeding passages [8]. Further discussion on (5) can be found in [9].

Under assumption (5), (4) can be simplified by using rescaled variables [9]:

$$\tilde{t} = \frac{\chi\mu}{D}t, \quad \tilde{x} = \frac{\sqrt{\chi\mu}}{D}x, \quad \tilde{v} = \text{sign}(\chi)\sqrt{\frac{\chi}{\mu}}v, \quad \tilde{u} = u. \tag{6}$$

This simplifies (4) to

$$\begin{cases} v_t + (u - \varepsilon_2 v^2)_x = \varepsilon_1 v_{xx}, \\ u_t + (uv)_x = u_{xx}, \end{cases} \quad x \in \mathbb{R}, \quad t > 0 \tag{7}$$

after dropping the tilde accent. Here the new parameters are

$$\varepsilon_1 = \frac{\varepsilon}{D} \geq 0, \quad \varepsilon_2 = \frac{\varepsilon}{\chi}. \tag{8}$$

The initial condition for (7) is

$$(v, u)(x, 0) = (v_0, u_0)(x), \quad x \in \mathbb{R}. \tag{9}$$

For a general background on (1), (7) and related models, readers are referred to [5, 7, 9] and references therein. Here we focus on results directly related to this paper. Equation (7) is a system of hyperbolic-parabolic conservation laws. If Cauchy data are small perturbations of a constant state $(0, \bar{u})$ with $\bar{u} > 0$, the L^2 theory of (7), (9) is well understood. This includes local existence, global existence, asymptotic decay rates, and convergence to an asymptotic solution, as a direct application of Kawashima’s theory [2, 3]. Similarly, pointwise estimates hence L^p theory with $p \geq 1$ are available, also as an application of the general theory [6].

If (v_0, u_0) is prescribed around $(0, \bar{u})$ but $(v_0, u_0 - \bar{u})$ has finite H^2 norm that is not necessary small, global existence of solution to (7), (9) has been established in [1, 5, 11] for the case $\varepsilon = 0$. In particular, under the additional zero-mass assumption on the perturbation and the smallness assumption on the initial perturbation and its anti-derivative, algebraic time decay rates in the L^2 framework are established in [5]. For the case $\varepsilon > 0$, similar results are obtained in [7].

The time decay rates in [5, 7] are obtained by energy and weighted energy methods. Although the rates are one can possibly have via those methods, they are not optimal. Here our purpose is to improve those rates to optimal ones through an iteration scheme based on spectral analysis, Green’s function and Duhamel’s principle. We also obtain corresponding rates for the original variables, i.e., the solution to (1), (2). In particular, we establish optimal rates of s and its derivatives in the border case $-\mu\bar{u} = \sigma$. This answers a question posted in [5], see Remark 1.2 therein. We further comment that similar results are obtained recently when (1) or (7) has a logistic growth term in the equation for cells [10].

Next we formulate the results from [5, 7], as they are the starting point of our analysis. We consider the scenario that (s_0, u_0) in (2) is prescribed around a constant state (\bar{s}, \bar{u}) , where \bar{s} and \bar{u} are positive. Correspondingly, from (3) and (6) we have (v_0, u_0) in (9) as prescribed around $(0, \bar{u})$. From (7), both v and $u - \bar{u}$ are conserved quantities. In particular, from (3) and (6),

$$\int_{\mathbb{R}} v(x, t) dx = \int_{\mathbb{R}} v_0(x) dx = \text{sign}(\chi) \sqrt{\frac{\chi}{\mu}} \int_{\mathbb{R}} \frac{d}{dx} (\ln s_0(x)) dx = 0.$$

If we assume $\int_{\mathbb{R}} [u_0(x) - \bar{u}] dx = 0$, we also have

$$\int_{\mathbb{R}} [u(x, t) - \bar{u}] dx = 0.$$

These allow us to define anti-derivatives:

$$\psi(x, t) \equiv \int_{-\infty}^x v(y, t) dy, \quad \phi(x, t) \equiv \int_{-\infty}^x [u(y, t) - \bar{u}] dy. \tag{10}$$

$$\psi_0(x) \equiv \psi(x, 0) = \int_{-\infty}^x v_0(y) dy, \quad \phi_0(x) \equiv \phi(x, 0) = \int_{-\infty}^x [u_0(y) - \bar{u}] dy. \tag{11}$$

We introduce some notations. Throughout this paper we use C to denote a universal positive constant, depending only on the system parameters and initial data. We also use the following notations to abbreviate the norms of Sobolev spaces with respect to x :

$$\|\cdot\|_k = \|\cdot\|_{H^k(\mathbb{R})}, \quad \|\cdot\| = \|\cdot\|_{L^2(\mathbb{R})}.$$

Theorem 1 ([5]) *Suppose that $u_0 \geq 0, \bar{u} > 0, (\psi_0, \phi_0) \in H^3(\mathbb{R})$ and there exists a sufficiently small constant $\eta_0 > 0$ such that $\|\psi_0\|_1^2 + \|\phi_0\|^2 \leq \eta_0$. Then there exists a unique global solution to (7)–(9) with $\varepsilon = 0$, satisfying $v \in C([0, \infty); H^2(\mathbb{R})) \cap L^2([0, \infty); H^2(\mathbb{R}))$ and $u - \bar{u} \in C([0, \infty); H^2(\mathbb{R})) \cap L^2([0, \infty); H^3(\mathbb{R}))$. Moreover, the solution has the decay estimate:*

$$\begin{aligned} & \sum_{k=0}^2 (t+1)^{k+1} \|D_x^k(v, u - \bar{u})\|^2(t) + \sum_{k=1}^2 \int_0^t (\tau+1)^k \|D_x^k v\|^2(\tau) d\tau \\ & + \sum_{k=1}^3 \int_0^t (\tau+1)^k \|D_x^k u\|^2(\tau) d\tau \leq C \quad t > 0. \end{aligned} \tag{12}$$

We comment that the statement of Theorem 1 is slightly different from Theorem 1.3 of [5]. This can be justified by a simple iteration, using Theorem 1.1 in [5]. See a similar argument for the model with logistic growth in [10].

Theorem 2 ([7]) *Suppose that $u_0 \geq 0, \bar{u} > 0, (\psi_0, \phi_0) \in H^3(\mathbb{R})$ and there exists a sufficiently small constant $\eta_0 > 0$ such that $\|(\psi_0, \phi_0)\|^2 \leq \eta_0$. Then there exists a unique global solution to (7)–(9) with $\varepsilon > 0$, satisfying $(v, u - \bar{u}) \in C([0, \infty); H^2(\mathbb{R})) \cap L^2([0, \infty); H^3(\mathbb{R}))$. Moreover, the solution has the decay estimate: For $t > 0$,*

$$\sum_{k=0}^2 (t+1)^{k+1} \|D_x^k(v, u - \bar{u})\|^2(t) + \sum_{k=0}^3 \int_0^t (\tau+1)^k \|D_x^k(v, u - \bar{u})\|^2(\tau) d\tau \leq C. \tag{13}$$

Our main results are the following theorems. The first one improves the L^2 decay rates of $(v, u - \bar{u})$ and its derivatives in (12) and (13) to optimal ones. The second one concerns the original variables s and u , or the solution to (1), (2).

Theorem 3 *Assume that $u_0 \geq 0, \bar{u} > 0$, and $(\psi_0, \phi_0) \in H^3(\mathbb{R}) \cap L^1(\mathbb{R})$.*

- *There exists a sufficiently small constant $\eta_0 > 0$ such that if $\|\psi_0\|_1^2 + \|\phi_0\|^2 \leq \eta_0$, the unique global solution to (7)–(9) with $\varepsilon = 0$, given in Theorem 1, satisfies*

$$\sum_{k=0}^1 (t + 1)^{\frac{3}{4} + \frac{k}{2}} \|D_x^k(v, u - \bar{u})\|(t) \leq C, \quad t > 0. \tag{14}$$

- *There exists a sufficiently small constant $\eta_0 > 0$ such that if $\|(\psi_0, \phi_0)\|^2 \leq \eta_0$, the unique global solution to (7)–(9) with $\varepsilon > 0$, given in Theorem 2, satisfies*

$$\sum_{k=0}^2 (t + 1)^{\frac{3}{4} + \frac{k}{2}} \|D_x^k(v, u - \bar{u})\|(t) \leq C, \quad t > 0. \tag{15}$$

Theorem 4 *Assume that $s_0 > 0, \bar{s} > 0, u_0 \geq 0, \bar{u} > 0$, and ϕ_0 be defined in (11). Let $(s_0 - \bar{s}, \phi_0) \in H^3(\mathbb{R}) \cap L^1(\mathbb{R})$. Then there exists a sufficiently small constant $\eta_0 > 0$ such that if $\|s_0 - \bar{s}\|_1^2 + \|\phi_0\|^2 \leq \eta_0$, the Cauchy problem (1), (2) with $\varepsilon \geq 0$ has a unique classical solution for $t \geq 0$, satisfying $s(x, t) > 0$ and $u(x, t) \geq 0$. We write*

$$s(x, t) = e^{-(\mu\bar{u} + \sigma)t} \tilde{s}(x, t). \tag{16}$$

Then the solution has the decay property for $t > 0$ as follows: If $\varepsilon = 0$,

$$\sum_{k=0}^2 (t + 1)^{\frac{1}{4} + \frac{k}{2}} \|D_x^k(\tilde{s} - \bar{s})\|(t) + \sum_{k=0}^1 (t + 1)^{\frac{3}{4} + \frac{k}{2}} \|D_x^k(u - \bar{u})\|(t) \leq C. \tag{17}$$

If $\varepsilon > 0$,

$$\sum_{k=0}^3 (t + 1)^{\frac{1}{4} + \frac{k}{2}} \|D_x^k(\tilde{s} - \bar{s})\|(t) + \sum_{k=0}^2 (t + 1)^{\frac{3}{4} + \frac{k}{2}} \|D_x^k(u - \bar{u})\|(t) \leq C. \tag{18}$$

We prove Theorem 3 in Sect. 2, and Theorem 4 in Sect. 3.

2 Decay Rates for the Transformed System

We write (7) in terms of the perturbation. Let

$$\begin{aligned} \tilde{u} &= u - \bar{u}, \quad \tilde{u}_0 = u_0 - \bar{u}, \\ w(x, t) &= \begin{pmatrix} v \\ \tilde{u} \end{pmatrix} (x, t) = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} (x, t), \quad \Phi(x, t) = \int_{-\infty}^x w(y, t) dy = \begin{pmatrix} \psi \\ \phi \end{pmatrix} (x, t), \end{aligned} \tag{19}$$

$$w_0(x) = \begin{pmatrix} v_0 \\ \tilde{u}_0 \end{pmatrix} (x) = \begin{pmatrix} w_{01} \\ w_{02} \end{pmatrix} (x), \quad \Phi_0(x) = \Phi(x, 0) = \begin{pmatrix} \psi_0 \\ \phi_0 \end{pmatrix} (x). \tag{20}$$

Then (7), (9) can be written as

$$\begin{cases} w_t + Aw_x = Bw_{xx} + R \\ w(x, 0) = w_0(x) \end{cases}, \tag{21}$$

$$A = \begin{pmatrix} 0 & 1 \\ \bar{u} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R = \tilde{R}_x, \quad \tilde{R} = \begin{pmatrix} \varepsilon_2 w_1^2 \\ -w_1 w_2 \end{pmatrix}. \tag{22}$$

Denote the Fourier transform of $w(x, t)$ with respect to x as $\hat{w}(\xi, t)$, etc. Then taking Fourier transform of (21) gives us

$$\begin{aligned} \hat{w}_t &= E(i\xi)\hat{w} + \hat{R}, \\ E(i\xi) &= -i\xi A - \xi^2 B. \end{aligned} \tag{23}$$

The solution of (23) is

$$\hat{w}(\xi, t) = e^{tE(i\xi)}\hat{w}(\xi, 0) + \int_0^t e^{(t-\tau)E(i\xi)}\hat{R}(\xi, \tau) d\tau. \tag{24}$$

To study the solution operator in (24), we perform spectral analysis for

$$E(i\xi) = \begin{pmatrix} -\varepsilon_1 \xi^2 & -i\xi \\ -\bar{u}i\xi & -\xi^2 \end{pmatrix} = \lambda_1(i\xi)P_1(i\xi) + \lambda_2(i\xi)P_2(i\xi),$$

where by direct calculation, the eigenvalues are

$$\lambda_{1,2}(i\xi) = -\frac{1}{2}(\varepsilon_1 + 1)\xi^2 \pm \sqrt{\frac{1}{4}(\varepsilon_1 + 1)^2\xi^4 - \xi^2(\varepsilon_1\xi^2 + \bar{u})}, \tag{25}$$

and the corresponding eigenprojections are

$$P_{1,2}(i\xi) = \frac{1}{-\bar{u}\xi^2 + (\lambda_{1,2} + \varepsilon_1\xi^2)^2} \begin{pmatrix} -\bar{u}\xi^2 & -i\xi(\lambda_{1,2} + \varepsilon_1\xi^2) \\ -\bar{u}i\xi(\lambda_{1,2} + \varepsilon_1\xi^2) & (\lambda_{1,2} + \varepsilon_1\xi^2)^2 \end{pmatrix}. \tag{26}$$

The solution operator in (24) is

$$e^{tE(i\xi)} = e^{\lambda_1(i\xi)t} P_1(i\xi) + e^{\lambda_2(i\xi)t} P_2(i\xi). \tag{27}$$

Noting that time decay rates are mainly determined by the behavior of solution operator for small ξ , we take Taylor expansions in (25) and (26): For $|\xi| \ll 1$,

$$\begin{aligned} \lambda_{1,2}(i\xi) &= -\frac{1}{2}(\varepsilon_1 + 1)\xi^2 \pm i\xi g(\xi), \quad g(\xi) = \sqrt{\bar{u}} + O(\xi^2) \in \mathbb{R}, \\ P_{1,2}(i\xi) &= \frac{1}{2} \begin{pmatrix} 1 + O(\xi) & \mp \frac{1}{\sqrt{\bar{u}}} + O(\xi^2) \\ \mp \sqrt{\bar{u}} + O(\xi^2) & 1 + O(\xi) \end{pmatrix}. \end{aligned} \tag{28}$$

We also need an estimate on the solution operator for $\xi \in \mathbb{R}$:

Lemma 1 *The solution operator in (24) satisfies*

$$|e^{tE(i\xi)}| \leq C e^{-\frac{c\xi^2 t}{1+\xi^2}}, \quad \xi \in \mathbb{R}, \quad t \geq 0, \tag{29}$$

where C and c are two positive constants depending on $\varepsilon_1 \geq 0$ and $\bar{u} > 0$ only.

Lemma 1 is an application of Kawashima’s theory [2]. A discussion of it can be found in [10], where a direct proof of Lemma 1 is also given. With (28) and Lemma 1, we are ready to prove the following decay estimate on the flow:

Lemma 2 *Let $\varepsilon_1 \geq 0$, $k \geq 0$ be an integer, $h = (h_1, h_2)^t \in L^1(\mathbb{R})$, $D_x^k h \in L^2(\mathbb{R})$. Then*

$$\|e^{tE(i\xi)}(i\xi)^k \hat{h}(\xi)\| \leq C(t+1)^{-\frac{1}{4}-\frac{k}{2}}(\|h_1\|_{L^1} + \|h_2\|_{L^1}) + C e^{-ct} \|D_x^k h\|, \quad t \geq 0, \tag{30}$$

where C and c are positive constants depending only on $\varepsilon_1 \geq 0$ and $\bar{u} > 0$.

Proof Let $\eta > 0$ be small such that (28) holds for $|\xi| \leq \eta$. We write

$$I \equiv \|e^{tE(i\xi)}(i\xi)^k \hat{h}(\xi)\|^2 = \left(\int_{|\xi| \leq \eta} + \int_{|\xi| \geq \eta} \right) |e^{tE(i\xi)}(i\xi)^k \hat{h}(\xi)|^2 d\xi.$$

Applying (27) and (28) to the first integral and (29) to the second one, we have

$$\begin{aligned} I &\leq \int_{|\xi| \leq \eta} C|\xi|^{2k} e^{-\xi^2 t} |\hat{h}(\xi)|^2 d\xi + \int_{|\xi| \geq \eta} C e^{-\frac{2c\eta^2 t}{1+\eta^2}} |(i\xi)^k \hat{h}(\xi)|^2 d\xi \\ &\leq C(t+1)^{-k-\frac{1}{2}} \|\hat{h}\|_{L^\infty}^2 + C e^{-\tilde{c}t} \|(i\xi)^k \hat{h}\|^2 \leq C(t+1)^{-k-\frac{1}{2}} \|h\|_{L^1}^2 + C e^{-\tilde{c}t} \|D_x^k h\|^2, \end{aligned}$$

where $\tilde{c} > 0$ is a constant, and we have used Plancherel theorem. Taking the square root we obtain (30). □

To prove Theorem 3, we only need to prove (14) for $\varepsilon = 0$ while (12) is valid, and prove (15) for $\varepsilon > 0$ while (13) is true. For this we use (21), which is equivalent to (7), (9). We consider $\varepsilon \geq 0$. By Plancherel theorem, (24) and the triangle inequality, for an integer $k \geq 0$ we have

$$\begin{aligned} \|D_x^k w\|(t) &= \|(i\xi)^k \hat{w}\|(t) \leq \|(i\xi)^k e^{tE(i\xi)} \hat{w}(\xi, 0)\| \\ &\quad + \int_0^t \|(i\xi)^k e^{(t-\tau)E(i\xi)} \hat{R}(\xi, \tau)\| d\tau \equiv I_1 + I_2. \end{aligned} \tag{31}$$

Noting (11) and (20), we have $w_0(x) = (\psi'_0, \phi'_0)^t(x)$. Thus applying (30) gives us

$$I_1 = \|(i\xi)^{k+1} e^{tE(i\xi)} (\hat{\psi}_0, \hat{\phi}_0)^t\| \leq C[(t+1)^{-\frac{3}{4}-\frac{k}{2}} (\|\psi_0\|_{L^1} + \|\phi_0\|_{L^1}) + Ce^{-ct} \|D_x^{k+1}(\psi_0, \phi_0)^t\|] \leq C(t+1)^{-\frac{3}{4}-\frac{k}{2}}, \quad 0 \leq k \leq 2. \tag{32}$$

Similarly, with (22) we have

$$I_2 \leq \int_0^t [C(t-\tau+1)^{-\frac{3}{4}-\frac{k}{2}} (\|w_1^2\|_{L^1} + \|w_1 w_2\|_{L^1})(\tau) + Ce^{-c(t-\tau)} (\|D_x^{k+1}(w_1^2)\| + \|D_x^{k+1}(w_1 w_2)\|)(\tau)] d\tau. \tag{33}$$

For the case $k = 0$, we define

$$M(t) = \sup_{0 \leq \tau \leq t} [(\tau+1)^{\frac{3}{4}} \|w\|(\tau)], \tag{34}$$

which implies $\|w\|(t) \leq M(t)(t+1)^{-\frac{3}{4}}$ for $t \geq 0$. With (12) and (13) we have

$$\begin{aligned} (\|w_1^2\|_{L^1} + \|w_1 w_2\|_{L^1})(\tau) &\leq (\|w_1\|^2 + \|w_1\| \|w_2\|)(\tau) \\ &= \|w_1\|^{\frac{1}{2}}(\tau) (\|w_1\|^{\frac{3}{2}} + \|w_1\|^{\frac{1}{2}} \|w_2\|)(\tau) \leq CM(\tau)^{\frac{1}{2}} (\tau+1)^{-\frac{9}{8}}. \end{aligned} \tag{35}$$

By Sobolev inequality, (12) and (13), we also have

$$\begin{aligned} (\|D_x(w_1^2)\| + \|D_x(w_1 w_2)\|)(\tau) &\leq C(\|w\|_{L^\infty} \|w_x\|)(\tau) \\ &\leq C(\|w\|^{\frac{1}{2}} \|w_x\|^{\frac{3}{2}})(\tau) \leq C(\tau+1)^{-\frac{7}{4}}. \end{aligned} \tag{36}$$

Substituting (35) and (36) into (33), for $k = 0$ we have

$$\begin{aligned} I_2 &\leq C \int_0^t [M(\tau)^{\frac{1}{2}} (t-\tau+1)^{-\frac{3}{4}} (\tau+1)^{-\frac{9}{8}} + e^{-c(t-\tau)} (\tau+1)^{-\frac{7}{4}}] d\tau \\ &\leq C[M(t)^{\frac{1}{2}} (t+1)^{-\frac{3}{4}} + (t+1)^{-\frac{7}{4}}]. \end{aligned} \tag{37}$$

Substituting (32) and (37) into (31) with $k = 0$, we have

$$\|w\|(t) \leq C(t+1)^{-\frac{3}{4}} + CM(t)^{\frac{1}{2}} (t+1)^{-\frac{3}{4}}.$$

Thus by (34) and Young inequality,

$$M(t) \leq C + CM(t)^{\frac{1}{2}} \leq C + \frac{1}{2}M(t),$$

which implies $M(t) \leq C$, hence

$$(t + 1)^{\frac{3}{4}} \|w\|(t) \leq C, \quad t \geq 0. \tag{38}$$

The case $k = 1$ is simpler as we are able to use the updated estimate (38) in (35) to give $(\|w_1^2\|_{L^1} + \|w_1 w_2\|_{L^1})(\tau) \leq C(\tau + 1)^{-\frac{3}{2}}$. Thus for this case,

$$I_2 \leq C \int_0^t [(t - \tau + 1)^{-\frac{5}{4}} (\tau + 1)^{-\frac{3}{2}} + e^{-c(t-\tau)} (\tau + 1)^{-\frac{9}{4}}] d\tau \leq C(t + 1)^{-\frac{5}{4}}. \tag{39}$$

Substituting (32) and (39) into (31) gives us $(t + 1)^{\frac{5}{4}} \|D_x w\|(t) \leq C$.

We only need to justify the term $k = 2$ in (15), which is for $\varepsilon > 0$. In this case, we replace (33) by

$$\begin{aligned} I_2 &\leq \int_0^{\frac{t}{2}} C(t - \tau + 1)^{-\frac{7}{4}} (\|w_1^2\|_{L^1} + \|w_1 w_2\|_{L^1})(\tau) d\tau \\ &\quad + \int_{\frac{t}{2}}^t C(t - \tau + 1)^{-\frac{5}{4}} (\|D_x(w_1^2)\|_{L^1} + \|D_x(w_1 w_2)\|_{L^1})(\tau) d\tau \\ &\quad + \int_0^t C e^{-c(t-\tau)} (\|D_x^3(w_1^2)\| + \|D_x^3(w_1 w_2)\|)(\tau) d\tau. \end{aligned}$$

With the updated estimates on $\|D_x^k w\|$, $k = 0, 1$, we have

$$\begin{aligned} I_2 &\leq C \int_0^{\frac{t}{2}} (t - \tau + 1)^{-\frac{7}{4}} (\tau + 1)^{-\frac{3}{2}} d\tau + C \int_{\frac{t}{2}}^t (t - \tau + 1)^{-\frac{5}{4}} (\tau + 1)^{-2} d\tau \\ &\quad + C \int_0^t e^{-c(t-\tau)} [(\tau + 1)^{-\frac{23}{8}} + (\tau + 1)^{-1} \|D_x^3 w\|(\tau)] d\tau \tag{40} \\ &\leq C(t + 1)^{-\frac{7}{4}} + C \left[\int_0^t e^{-2c(t-\tau)} (\tau + 1)^{-5} d\tau \right]^{\frac{1}{2}} \left[\int_0^t (\tau + 1)^3 \|D_x^3 w\|^2(\tau) d\tau \right]^{\frac{1}{2}} \\ &\leq C(t + 1)^{-\frac{7}{4}}, \end{aligned}$$

where we have used Cauchy-Schwarz inequality and (13). Now combining (31), (32) and (40) gives $(t + 1)^{\frac{7}{4}} \|D_x^2 w\|(t) \leq C$. Thus we have proved Theorem 3.

The following is a natural extension of this section, and is needed in next section. Using notations in (19), (20) and (22), we integrate (21) to have

$$\Phi_t + A\Phi_x = B\Phi_{xx} + \tilde{R}.$$

Thus similar to (24),

$$\hat{\Phi}(\xi, t) = e^{tE(i\xi)} \hat{\Phi}(\xi, 0) + \int_0^t e^{(t-\tau)E(i\xi)} \hat{R}(\xi, \tau) d\tau.$$

Following (31) – (33) and applying Theorem 3, we have

$$\begin{aligned}
 \|\Phi\|(t) &= \|\hat{\Phi}\|(t) \leq \|e^{tE(i\xi)}\hat{\Phi}_0\| + \int_0^t \|e^{(t-\tau)E(i\xi)}\hat{R}(\xi, \tau)\| d\tau \\
 &\leq C(t+1)^{-\frac{1}{4}} + C \int_0^t [(t-\tau+1)^{-\frac{1}{4}}(\|w_1^2\|_{L^1} + \|w_1w_2\|_{L^1})(\tau) \\
 &\quad + e^{-c(t-\tau)}(\|w_1^2\| + \|w_1w_2\|)(\tau)] d\tau \\
 &\leq C(t+1)^{-\frac{1}{4}}.
 \end{aligned}
 \tag{41}$$

3 Decay Rates for the Original System

To simplify our notations and without loss of generality, we assume $\tilde{t} = t$, $\tilde{x} = x$, and $\tilde{v} = v$ in (6). To prove Theorem 4 we first note that under the hypotheses of the theorem, the assumptions in Theorem 3 are satisfied for each of the cases $\varepsilon = 0$ and $\varepsilon > 0$. This is in view of (3) and (11), which imply $\psi_0(x) = \ln s_0(x) - \ln \bar{s}$, hence $|\psi_0(x)| \leq \frac{2}{\bar{s}}|s_0(x) - \bar{s}|$ and $|\psi'_0(x)| \leq \frac{2}{\bar{s}}|s'_0(x)|$ for small $\|s_0 - \bar{s}\|_1$. Thus (7)–(9) has a unique global solution, satisfying (14) and (15) for $\varepsilon = 0$ and $\varepsilon > 0$, respectively. The inverse transform of (3),

$$s(x, t) = e^{-(\mu\bar{u}+\sigma)t}\tilde{s}(x, t), \quad \tilde{s}(x, t) = \bar{s}e^{\psi(x,t)}, \tag{42}$$

then gives us a unique, global solution to (1), (2).

The inverse transform (42) implies $s(x, t) > 0$ for all $x \in \mathbb{R}$ and $t \geq 0$. Applying the maximum principle to the second equation in (7), one concludes that $u(x, t) \geq 0$ as well, provided $u_0(x) \geq 0$. A similar, detailed discussion can be found in [9] for the model with logistic growth. As the estimates for $u - \bar{u}$ in (17) and (18) are inherited from (14) and (15), respectively, we obtain those for $\tilde{s} - \bar{s}$ below.

From (19) and (41), we have

$$\|\psi\|(t) \leq C(t+1)^{-\frac{1}{4}}. \tag{43}$$

Since $\psi_x = v$, by Sobolev inequality, (14) and (15), we further have

$$\|\psi\|_{L^\infty}(t) \leq C\|\psi\|^{\frac{1}{2}}(t)\|v\|^{\frac{1}{2}}(t) \leq C(t+1)^{-\frac{1}{2}}. \tag{44}$$

Therefore,

$$\|\tilde{s}\|_{L^\infty}(t) \leq \bar{s}e^{\|\psi\|_{L^\infty}(t)} \leq C. \tag{45}$$

From (42), (45) and the mean value theorem, we have

$$\begin{aligned}
 |\tilde{s}(x, t) - \bar{s}| &= \bar{s}|e^{\psi(x,t)} - 1| \leq \bar{s}e^{\|\psi\|_{L^\infty}(t)}|\psi(x, t)| \leq C|\psi(x, t)|, \\
 \tilde{s}_x(x, t) &= (\tilde{s}v)(x, t), \quad \tilde{s}_{xx}(x, t) = (\tilde{s}v^2 + \tilde{s}v_x)(x, t), \\
 \tilde{s}_{xxx}(x, t) &= (\tilde{s}v^3 + 3\tilde{s}vv_x + \tilde{s}v_{xx})(x, t).
 \end{aligned}$$

Together with (43), (45), (14) and (15), these give us

$$\begin{aligned} \|\tilde{s} - \bar{s}\|(t) &\leq C\|\psi\|(t) \leq C(t+1)^{-\frac{1}{4}}, \quad \|\tilde{s}_x\|(t) \leq (\|\tilde{s}\|_{L^\infty}\|v\|)(t) \leq C(t+1)^{-\frac{3}{4}}, \\ \|\tilde{s}_{xx}\|(t) &\leq \|\tilde{s}\|_{L^\infty}(t)(\|v\|_{L^\infty}\|v\| + \|v_x\|)(t) \leq C(t+1)^{-\frac{5}{4}}. \end{aligned}$$

In the case $\varepsilon > 0$ we also have

$$\|D_x^3\tilde{s}\|(t) \leq C\|\tilde{s}\|_{L^\infty}(t)(\|v\|^2\|v_x\| + \|v\|^{\frac{1}{2}}\|v_x\|^{\frac{3}{2}} + \|v_{xx}\|)(t) \leq C(t+1)^{-\frac{7}{4}}.$$

We thus settle (17) and (18).

Acknowledgements Y. Zeng was partially supported by the National Science Foundation under grant DMS-1908195. K. Zhao was partially supported by the Simons Foundation Collaboration Grant for Mathematicians No. 413028.

References

1. Guo, J., Xiao, J., Zhao, H., Zhu, C.: Global solutions to a hyperbolic-parabolic coupled system with large initial data. *Acta Math. Sci. Ser. B* **29**, 629–641 (2009)
2. Kawashima, S.: Systems of a hyperbolic-parabolic composite type, with applications to the equations of magnetohydrodynamics. Doctoral thesis, Kyoto University (1983)
3. Kawashima, S.: Large-time behaviour of solutions to hyperbolic-parabolic systems of conservation laws and applications. *Proc. Roy. Soc. Edinburgh Sect. A* **106**, 169–194 (1987)
4. Levine, H.A., Sleeman, B.D.: A system of reaction diffusion equations arising in the theory of reinforced random walks. *SIAM J. Appl. Math.* **57**, 683–730 (1997)
5. Li, D., Pan, R., Zhao, K.: Quantitative decay of a one-dimensional hybrid chemotaxis model with large data. *Nonlinearity* **28**, 2181–2210 (2015)
6. Liu, T.-P., Zeng, Y.: Large time behavior of solutions for general quasilinear hyperbolic-parabolic systems of conservation laws. *Mem. Amer. Math. Soc.* **125**, no. 599, viii+120 pp. (1997)
7. Martinez, V.R., Wang, Z., Zhao, K.: Asymptotic and viscous stability of large-amplitude solutions of a hyperbolic system arising from biology. *Indiana Univ. Math. J.* **67**, 1383–1424 (2018)
8. Othmer, H., Stevens, A.: Aggregation, blowup and collapse: the ABC's of taxis in reinforced random walks. *SIAM J. Appl. Math.* **57**, 1044–1081 (1997)
9. Zeng, Y., Zhao, K.: On the logarithmic Keller-Segel-Fisher/KPP system. *Disc. Cont. Dyn. Syst. Ser. A* **39**, 5365–5402 (2019)
10. Zeng, Y., Zhao, K.: Optimal decay rates for a chemotaxis model with logistic growth, logarithmic sensitivity and density-dependent production/consumption rate. *J. Differ. Equ.* **268**, 1379–1411 (2020)
11. Zhang, Y., Tan, Z., Sun, M.-B.: Global existence and asymptotic behavior of smooth solutions to a coupled hyperbolic-parabolic system. *Nonlinear Anal. Real World Appl.* **14**, 465–482 (2013)

Mathematical and Statistical Modelling in Life Sciences

An Optimal Control Strategy for a Malaria Model



Onoja Abu and Ikechukwu Ignatius Ayogu

Abstract Malaria is a major vector-borne disease that has been generating a serious health burden and devastating the economy of Sub-Saharan Africa, South-East Asia, the Eastern Mediterranean, Western Pacific and Americas. In this paper, a mathematical model for low and high malaria risk human population groups, incorporating four control variables representing insecticide treated nets, treatment, indoor residual spraying and intermittent preventive treatment; seasonally forced mosquito population and transmission parameters, is formulated. The necessary conditions for the optimality of the model are derived using the Pontryagin's Maximum Principle. The optimal control model is numerically explored using Runge-Kutta method of order four. Experimental results show that the model is able to indicate the best control strategy, given the estimated costs of implementation of the varying control measures.

Keywords Optimal control strategies · Malaria disease · Cost-effectiveness · Plasmodium species

1 Introduction

Malaria parasites are amongst organisms that live in other organisms as host-dependent guests [1]. Malaria is a disease caused by infection with protozoan parasites belonging to the genus Plasmodium transmitted by infected female Anopheles mosquitoes through bites when taking blood meal [2, 3]. The four species that commonly infect humans are: Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae.

O. Abu (✉)

Department of Mathematics and Statistics, The Federal Polytechnic, Idah Kogi, Nigeria

e-mail: abuonoja2008@yahoo.com

I. Ignatius Ayogu

Department of Computer Science, The Federal Polytechnic, Idah Kogi, Nigeria

e-mail: ig.ayogu@ieee.org

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_43

465

Malaria has caused huge health and financial burdens. Although, the number of malaria cases globally dropped from an estimated 262 million in 2000 to 219 million cases in 2017, the burden is still very significant, especially in the worst afflicted WHO African Region where 200 million or 92% of the cases and 93% of malaria related deaths occurred in 2017 [5].

Malaria poses serious financial and economic burdens to governments and households of malarious countries. Estimates of US\$ 3.1, US\$ 2.7, US\$ 2.9, US\$ 2.5 and US\$ 2.7 billion were invested in malaria control and elimination efforts in 2017, 2016, 2015, 2014 and 2013 respectively across the globe by governments of malaria endemic countries and international partners [3–7]. Global financing for malaria control increased from an estimated US\$ 960 million in 2005 to US\$ 2.7 billion in 2013 [4–8]. In Africa malaria affected the national income to the tune of 0.6–1.0% of its gross domestic product (GDP); in Kenya, up to 2–6% of her GDP, and at 1–5% for Nigeria [9].

To avert malaria health burden, governments of endemic countries and international donors have put some major preventive and control measures in place. These include use of insecticide-treated nets (ITNs), artemisinin based combination therapy (ACT), indoor residual spraying (IRS) and intermittent preventive treatment (IPT) [3–7]. Studies to evaluate the financial and economic costs or cost effectiveness analysis of these interventions in similar or different localities against similar or different health outcomes have been performed. For details, the reader is referred to [10–12].

Mathematical modeling has been an important tool to study many processes, including the dynamics of infectious diseases. We have reviewed some relevant malaria models suitable to our work. Cognizance is taken mostly of the ordinary differential equation models that either feature seasonality alone or incorporate control variables. Optimal control models for infectious diseases abound. Optimal control theory is a mathematical technique for steering a dynamical system. Optimal control techniques can be found in [13]. Malaria models with control variables can be seen in [14–25].

The goal of this paper is to formulate an optimal control model for malaria, analytically investigate the existence of optimal control vector and numerically explore the corresponding optimality system.

2 Formulation of the Optimal Control Model

We use the following tips as a guide in the formulation of our model. In the human population, some groups are more vulnerable than others. These include pregnant women, children below 5 years and people with immunity impairment such as HIV patients, immigrants or travelers from malaria-free areas. Malaria causes deaths, especially, in *Plasmodium falciparum* endemic settings. Malaria-related deaths have economic cost implications. Mosquito population and parasite development flourish

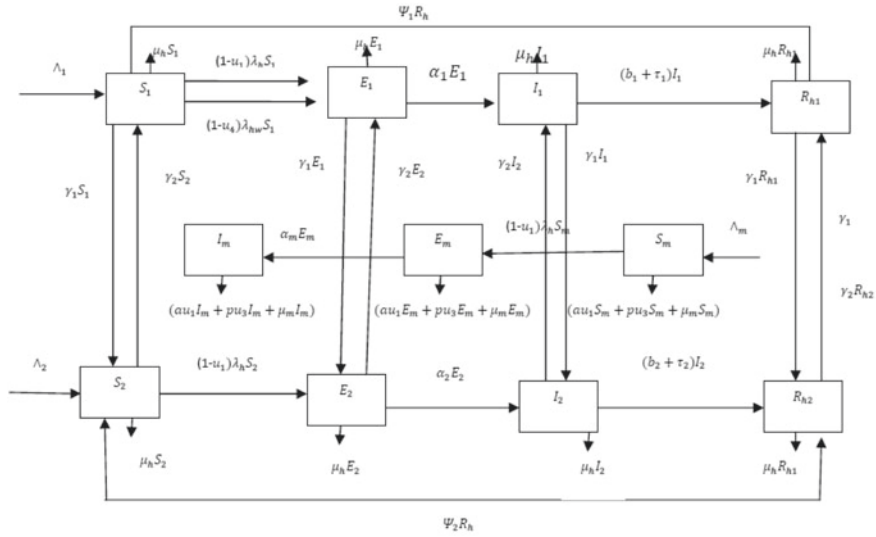


Fig. 1 Flow diagram of malaria transmission dynamics model

more in certain seasons than others; and therefore control efforts of malaria can be more effective and economical if they are in phases with seasonal variations. Malaria is endemic in many countries and parts of the world with seasonal changes. The control measures that are WHO recommended include ITNs, treatment, IRS and IPT for pregnant women, infants and children. ITNs, treatment, IRS and IPT are adopted by many malarious countries of the world. In this paper, we formulate an optimal control model

1. That classifies the host population into low and high risk groups;
2. Whose objective functional incorporates economic costs associated with malaria-induced death, exposed and infected humans, mosquitoes and the costs of implementation of controls;
3. That incorporates ITNs, treatment, IRS and IPT as controls;
4. That incorporates seasonally forced mosquito birth function and transmission parameter.

Optimal control models that contain all the four aforementioned components are rare to the best of our knowledge.

Figure 1 shows the flow of all the processes captured in the model while Tables 1, 2 and 3 describe all the variables and parameters used in the model.

Table 1 State variables for the malaria model

Variables	Description
$S_1(t)$	Population of high-risk susceptible individuals at time t
$E_1(t)$	Population of high-risk exposed individuals at time t
$I_1(t)$	Population of high-risk infectious humans at time t
$S_2(t)$	Population of low-risk susceptible individuals at time t
$E_2(t)$	Population of low-risk exposed individuals at time t
$I_2(t)$	Population of low-risk infectious humans at time t
$R_{h1}(t)$	Population of high-risk recovered humans at time t
$R_{h2}(t)$	Population of low-risk recovered humans at time t
$S_m(t)$	Population of susceptible mosquitoes at time t
$E_m(t)$	Population of exposed mosquitoes at time t
$I_m(t)$	Population of infectious mosquitoes at time t
$N_h(t)$	Total human population at time t
$N_{hw}(t)$	Total pregnant women population at time t
$N_m(t)$	Total mosquito population at time t

2.1 The Optimal Control Problem

The control problem is presented in the sequel. The cost or objective functional is given by

$$\begin{aligned}
 J(\mu_1, \mu_2, \mu_3, \mu_4) = & \int_0^t [A_1 E_1 + A_2 E_2 + A_3 I_1 + A_4 I_2 + A_5 N_m + A_6(\delta_1 I_1 + \delta_2 I_2) \\
 & + \frac{1}{2}(B_1 \mu_1^2 + B_2 \mu_2^2 + B_3 \mu_3^2 + B_4 \mu_4^2)] dt \tag{1}
 \end{aligned}$$

Subject to the state equations:

$$\frac{dS_1}{dt} = \Lambda_1 + \gamma_2 S_2 + \Psi_1 R_{h1} - (1 - \mu_1)\lambda_{h1} S_1 - (1 - r\mu_4)\lambda_{h1} S_1 - (\gamma_1 + \mu_h) S_1 \tag{2}$$

$$\frac{dE_1}{dt} = \gamma_2 E_2 + (1 - \mu_1)\lambda_{h1} S_1 + (1 - r\mu_4)\lambda_{h1} S_1 - (\gamma_1 + \alpha_1 + \mu_h) E_1 \tag{3}$$

$$\frac{dI_1}{dt} = \gamma_2 I_2 + \alpha_1 E_1 - (\delta_1 + \mu_h) I_1 - (\gamma_1 + b_1 + \tau_2 \mu_2) I_1 \tag{4}$$

Table 2 Parameters/variables of the malaria model

Param/Var	Description
φ	Mosquito contact rate with human
ε	Mosquito biting rate
ε_0	Transmission coefficient of infection from infectious mosquito to a high-risk susceptible human, provided there is a bite
ε_1	Transmission coefficient of infection from infectious mosquito to a low-risk susceptible human, provided there is a bite
λ	Transmission coefficient of infection from infectious human to a susceptible mosquito, provided there is a bite
δ_1	Per capita disease-induced mortality rate of high-risk infectious humans
δ_2	Per capita disease-induced mortality rate of low-risk infectious humans
μ_h	Per capita natural mortality rate of humans
μ_m	Per capita natural mortality rate of mosquitoes
\wedge_1	Recruitment rate into high-risk human population
\wedge_2	Recruitment rate into low-risk human population
λ_0	Recruitment of mosquitoes by birth (baseline)
Ψ_1	Per capita rate of loss of imunity of recovered individuals in high-risk group
Ψ_2	Per capita rate of loss of imunity of recovered individuals in low-risk group
α_1	Progression rate from high-risk, exposed to high-risk infected
α_2	Progression rate from low-risk, exposed to low-risk infected
b_1	Proportion of spontaneous recovery from high-risk population
b_2	Proportion of spontaneous recovery from low-risk population
λ_{hw}	Force of infection for susceptible pregnant women to exposed individuals
λ_m	Force of infection from susceptible mosquitoes to exposed mosquitoes
ω_0	Strength of seasonality
γ_1	Progression from high-risk group to low-risk group
γ_2	Progression from low-risk group to high-risk group

$$\frac{dS_2}{dt} = \wedge_2 + \gamma_1 S_1 + \Psi_2 R_{h2} - (1 - \mu_1)\lambda_{h2}S_2 - (\gamma_2 + \mu_h)S_2 \tag{5}$$

$$\frac{dE_2}{dt} = \gamma_1 E_1 + (1 - \mu_1)\lambda_{h2}S_2 - (\gamma_2 + \alpha_2 + \mu_h)E_2 \tag{6}$$

$$\frac{dI_2}{dt} = \gamma_1 I_1 + \alpha_2 E_2 - (\delta_2 + \mu_h)I_2 - (\gamma_2 + b_2 + \tau_2\mu_2)I_2 \tag{7}$$

$$\frac{dR_{h1}}{dt} = \gamma_2 R_{h2} + (b_1 + \tau_2\mu_2)I_1 - (\gamma_1 + \Psi_1 + \mu_h)R_{h1} \tag{8}$$

Table 3 Control variables/parameters in the model

Parameter	Description
$\mu_1(t)$	Insecticide-treated bed nets (ITN)
$\mu_2(t)$	Treatment of infectious individuals
$\mu_3(t)$	Indoor residual spraying (IRS)
$\mu_4(t)$	Intermittent prevent treatment for pregnant women
p	Efficacy of use of indoor residual spraying
τ	Efficacy of use of treatment
a	Efficacy of use of insecticide-treated bed nets
r	Efficacy for use of IPT
A_1	Cost associated with an exposed individual in high-risk population
A_2	Cost associated with an exposed individual in low-risk population
A_3	Cost associated with an infectious individual in high-risk population
A_4	Cost associated with an infectious individual in low-risk population
A_5	Cost associated with a mosquito
A_6	Cost associated with a human death
B_1	Cost of implementation of ITNs
B_2	Cost of implementation of treatment
B_3	Cost of implementation of IRS
B_4	Cost of implementation of IPT

$$\frac{dR_{h2}}{dt} = \gamma_2 R_{h1} + (b_2 + \tau_2 \mu_2) I_2 - (\gamma_2 + \Psi_2 + \mu_h) R_{h2} \tag{9}$$

$$\frac{dS_m}{dt} = \wedge_m - (1 - \mu_1) \lambda_m S_m - (\mu_m + a \mu_1 + p \mu_3) S_m \tag{10}$$

$$\frac{dE_m}{dt} = (1 - \mu_1) \lambda_m S_m - \alpha_m E_m - (\mu_m + a \mu_1 + p \mu_3) E_m \tag{11}$$

$$\frac{dI_m}{dt} = \alpha_m E_m - (\mu_m + a \mu_1 + p \mu_3) I_m \tag{12}$$

where: $\lambda_{h1} = \frac{\beta_1 \epsilon \Phi I_m}{N_h}$, $\lambda_{h2} = \frac{\beta_2 \epsilon \Phi I_m}{N_h}$, $\lambda_m = \frac{\lambda \epsilon \Phi I_1 + \lambda \epsilon \Phi I_2}{N_h}$, $\beta_1 = \zeta_0 (1 + \omega_0 \cos 2\pi t)$, $\beta_2 = \zeta_1 (1 + \omega_0 \cos 2\pi t)$, $\wedge_m = \lambda_0 (1 + \omega_0 \cos 2\pi t)$ and $\mu_1, \mu_2, \mu_3, \mu_4$ are Lebesgue measurable functions belonging to the set Ω . $\omega_0 \cos 2\pi t$ captures seasonal variation in transmission rate over time. This function is also a time translate of $\omega_0 \sin 2\pi t$.

The dynamics of the human host and the mosquito population are given by Eqs. (13) and (14)

$$\frac{dN_h}{dt} = \wedge_1 + \wedge_2 - \mu_h N_h - \delta_1 I_1 - \delta_2 I_2, \tag{13}$$

$$\frac{dN_m}{dt} = \wedge_m - \mu_m N_m. \tag{14}$$

2.2 Analysis of the Optimal Control Problem

The Lagrangian of the optimal control problem is the integrand of the objective functional and is given by Eq. (15)

$$\begin{aligned} L(I_1, I_2, E_1, E_2, N_m, \mu_1, \mu_2, \mu_3, \mu_4) = & A_1 E_1 + A_2 E_2 + A_3 I_1 + A_4 I_2 + A_5 N_m \\ & + A_6(\delta_1 I_1 + \delta_2 I_2) + \frac{1}{2}(B_1 \mu_1^2 + B_2 \mu_2^2 \\ & + B_3 \mu_3^2 + B_4 \mu_4^2). \end{aligned} \tag{15}$$

The Hamiltonian is given by $H = L + \sum_{i=1}^{11} \lambda_i f_i$, f_i 's are the right hand sides of Eqs. 1–12.

2.3 Optimality System

Suppose $U = (\mu_1, \mu_2, \mu_3, \mu_4)$ is a control vector, $x = (S_1, E_1, I_1, R_{h1}, S_2, E_2, R_{h2}, S_m, E_m, I_m)$ the state vector and H, the Hamiltonian, the optimality system is given by equation $\frac{dx_i}{dt} = \frac{\partial H}{\partial \lambda_i}$, $-\frac{d\lambda_i}{dt} = \frac{\partial H}{\partial x_i}$, $i = 1, \dots, 11$ with transversality conditions: $\lambda_i(tf) = 0$, $\frac{\partial H}{\partial \mu_j} = 0$, $j = 1, \dots, 4$.

Theorem 1 *Let $U = (\mu_1, \mu_2, \mu_3, \mu_4)$ be a control vector, $x = (S_1, E_1, I_1, R_{h1}, S_2, E_2, I_2, R_{h2}, S_m, E_m, I_m)$ be the state vector of the system (1–12) and H the Hamiltonian. There exist an optimal control vector $U^*(t)$ and the corresponding state vector $x^*(t)$ that minimize $J(U)$ over Ω . Furthermore, there exist adjoint functions λ_i satisfying the equations $-\frac{d\lambda_i}{dt} = \frac{\partial H}{\partial x_i}$ with transversality conditions $\lambda_i(tf) = 0$. In addition, the optimality controls are given by $u_j^* = \max\{0, \min(1, R_j)\}$, $j = 1, \dots, 4$.*

Proof We use the recipe by [13]. The existence of an optimal control vector follows from the convexity of the integrand J with respect to U , a priori boundedness of the state solutions and the Lipschitz property of the state solutions with respect to the state variables. See [13] (Corollary 4.1). The adjoint equations and transversality conditions can be obtained by using the Pontryagin’s Maximum Principle such that

$$\begin{aligned}
 &-\frac{d\lambda_1}{dt} = \frac{\partial H}{\partial S_1}; -\frac{d\lambda_2}{dt} = \frac{\partial H}{\partial E_1}; -\frac{d\lambda_3}{dt} = \frac{\partial H}{\partial I_1}; -\frac{d\lambda_4}{dt} = \frac{\partial H}{\partial S_2}; -\frac{d\lambda_5}{dt} = \frac{\partial H}{\partial E_1}; -\frac{d\lambda_6}{dt} = \frac{\partial H}{\partial I_2}; \\
 &-\frac{d\lambda_7}{dt} = \frac{\partial H}{\partial R_{h1}}; -\frac{d\lambda_8}{dt} = \frac{\partial H}{\partial R_{h2}}; -\frac{d\lambda_9}{dt} = \frac{\partial H}{\partial S_m}; -\frac{d\lambda_{10}}{dt} = \frac{\partial H}{\partial E_m}; -\frac{d\lambda_{11}}{dt} = \frac{\partial H}{\partial I_m};
 \end{aligned}$$

with transversality conditions

$$\lambda_1(T) = \lambda_2(T) = \lambda_3(T) = \lambda_4(T) = \lambda_5(T) = \lambda_6(T) = \lambda_7(T) = \lambda_8(T) = \lambda_9(T) = \lambda_{10}(T) = \lambda_{11}(T) = 0.$$

The controls u_j can be solved for by using the optimality conditions

$$-\frac{\partial H}{\partial u_1} = 0; -\frac{\partial H}{\partial u_2} = 0; -\frac{\partial H}{\partial u_3} = 0; -\frac{\partial H}{\partial u_4} = 0.$$

Therefore:

$$\begin{aligned}
 \mu_1^* &= \max\{0, \min(1, R_1)\}, \\
 R_1 &= \frac{\lambda_{h1}S_1(\lambda_2 - \lambda_1) + \lambda_{h2}S_2(\lambda_5 - \lambda_4) + \lambda_m S_m(\lambda_{10} - \lambda_9) + aS_m\lambda_9 + aE_m\lambda_{10} + aI_m\lambda_{11}}{B_1}
 \end{aligned} \tag{16}$$

$$\mu_2^* = \max\{0, \min(1, R_2)\}, R_2 = \frac{r_1(\lambda_3 - \lambda_7)I_1 + \tau_2(\lambda_6 - \lambda_8)I_2}{B_2} \tag{17}$$

$$\mu_3^* = \max\{0, \min(1, R_3)\}, R_3 = \frac{p(S_m + E_m\lambda_{10} + I_m\lambda_{11})}{B_3} \tag{18}$$

$$\mu_4^* = \max\{0, \min(1, R_4)\}, R_4 = \frac{(\lambda_2 - \lambda_1)\lambda_{h1}rS_1}{B_4} \tag{19}$$

3 Numerical Simulations and Results

For numerical simulation, we apply all the parameter values published in the literature and estimated others during the research process. In addition we use the following initial values and weight constants

$S_1(0) = 1450, E_1(0) = 250, I_1(0) = 205, R_{h1}(0) = 50, S_2(0) = 17000, E_2(0) = 125, I_2(0) = 125, R_{h2}(0) = 50, S_m(0) = 20000, E_m(0)5000, I_m(0) = 5000, A_1 = 1, A_2 = 1, A_3 = 20, A_4 = 16, A_5 = 0.1906, A_6 = 3000, B_1 = 24.10, B_2 = 10.64, B_3 = 73.42, B_4 = 12.21, \gamma_1 = 0.02, \varphi = 0.502, \epsilon = 0.4, \epsilon_0 = 0.0655, \epsilon_1 = 0.04, \lambda = 0.42, \delta_1 = \delta_2 = 0.05, \mu_h = 0.00004892, \mu_m = 0.04, \wedge_1 = 0.4202, \wedge_2 = 0.1, \lambda_0 = 2800, \psi_1 = \psi_2 = 0.01095, \gamma_1 = 0.02, r = 0.73, a = 0.51, \tau = 0.5, p = 0.51, u_1(t) = 0 - 1, u_2(t) = 0 - 1, u_3(t) = 0 - 1, u_4(t) = 0 - 1, \omega_0 = 0.7, \lambda_m = 0.00000048, \lambda_{hw} = 0.00000247, b_1 = 0.005, b_2 = 0.01, \alpha_m = 0.091, \alpha_1 = 0.1, \alpha_2 = 0.1.$

For simulation, we consider all the possible intervention strategies: one, two, three and four control strategies: ITN only, TRT only, IRS only, IPT only, ITN and TRT,

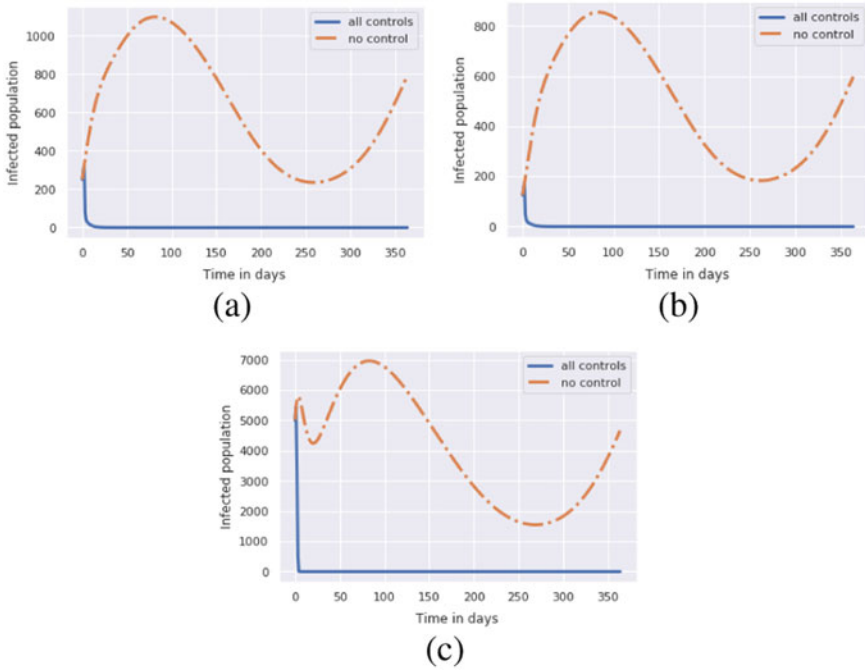


Fig. 2 Effects of all the control strategies on the population of **a** infected high-risk humans; **b** infected low-risk humans; **c** infected mosquitoes

ITN and IRS, ITN and IPT, TRT and IRS, TRT and IPT, IRS and IPT, ITN, TRT and IRS, ITN, TRT and IPT, ITN, IRS and IPT, TRT, IRS and IPT, ITN, TRT, IRS and IPT. The best numerical results are shown in Fig. 2a, b, c.

4 Discussion

This section discusses analytical and numerical results of our optimal control model. The optimal control model incorporates four time-dependent variables consisting of insecticide-treated bed nets (ITNs), treatment (TRT), indoor residual spraying (IRS) and intermittent preventive treatment (IPT) for high-risk humans. The main analytical result on existence and characterization of the corresponding optimality system of the control model can be found in Theorem 1. The optimality system was numerically explored for different possible intervention strategies. Table 4 shows the costs implications of all the different strategies while Fig. 2a, b, c shows the best outcomes of our numerical experiments. Table 4 shows that IPT as a control incurs the highest cost. The total cost of this intervention based on our model is \$10697856.67

Table 4 Cost of intervention strategies

Control level	Costs estimates	Control level	Costs estimates
No control	10884587.74	ITN/IPT	2378872.77
ITN	2442913.89	TRT/IRS	886414.14
TRT	6730937.95	TRT/IPT	6671496.03
IRS	2551439.20	IRS/IPT	2521914.07
IPT	10697856.67	ITN/TRT/IRS	343974.17
ITN/TRT	545385.27	ITN/TRT/IPT	544669.92
ITN/IRS	2180294.50	ITN/IRS/IPT	2169049.85
TRT/IRS/IPT	872797.00	ITN/TRT/IRS/IPT	343838.63

as against \$10884587.74 without control. In another development, ITN, Treatment, IRS and IPT as a strategy incurs the lowest cost of \$343838.63. The effects of this strategy on infected human and mosquito populations are depicted in Fig. 2a, b, c.

5 Conclusion

In this paper, we formulated an optimal control model for malaria, bearing in mind, the effects of seasonality on mosquito birth rate, the development of plasmodium parasites and the consequent implication on transmission parameters in an endemic setting with a seasonally forced mosquito population. Both the analytical and numerical results were obtained. The main analytical result on existence and optimality system can be found in Theorem 1. The costs of the different possible strategies can be seen in Table 4. The results show that optimal use of ITNs, IRS, TRT and IPT as a control strategy eliminates malaria fastest and gives the lowest cost. Optimal use of ITNs, IRS, TRT and IPT is the most cost-effective and therefore, recommended in an endemic setting where the mosquito population is seasonally forced.

References

1. Lucius, R., Poulin, R.: *Biology of Parasites*. Wiley (2016)
2. Cox, F.E.G.: History of the discovery of the malaria parasites and their vectors—a review. *Parasit. Vectors* **3**(5), 1–9 (2010)
3. WHO: *World Malaria Report 2015*, World Health Organization, WHO Global Malaria Programme, Geneva, Switzerland (2015)
4. WHO: *World Malaria Report 2018*, World Health Organization, WHO Global Malaria Programme, Geneva (2018)
5. WHO: *World Malaria Report 2017*, World Health Organization, WHO Global Malaria Programme, Geneva (2017)

6. WHO: World Malaria Report 2016, World Health Organization, WHO Global Malaria Programme, Geneva, Switzerland (2016)
7. WHO: World Malaria Report 2014, World Health Organization, WHO Global Malaria Programme, Geneva, Switzerland (2014)
8. WHO: World Malaria Report 2013, World Health Organization, WHO Global Malaria Programme, Geneva, Switzerland (2013)
9. WHO: World Malaria Report 1999, World Health Organization, WHO Global Malaria Programme, Geneva, Switzerland (1999)
10. White, M.T., Conteh, L., Cibulskis, R., Ghani, A.C.: Costs and cost-effectiveness of malaria control interventions—a systematic review. *Malar. J.* **10**, 337 (2011)
11. Fleming, W.H., Rishel, R.W.: *Deterministic and Stochastic Optimal Control*. Springer (2012)
12. Nana-Kyere, S., Doe, R.H., Boateng, F.A., Odum, J.K., Marmah, S., Banon, D.T.: Optimal control model of malaria disease with standard incidence rate. *J. Adv. Math. Comput. Sci.* **23**(5), 1–21 (2017)
13. Nemananzhe, L.: A mathematical modeling of optimal vaccination strategies in epidemiology. Master's thesis, Department of Mathematics and Applied Mathematics, University of the Western Cape (2010)
14. Mwanga, G.G., Haario, H., Nannyonga, B.K.: Optimal control of malaria model with drug resistance in presence of parameter uncertainty. *Appl. Math. Sci.* **8**(55), 2701–2730 (2014)
15. Mwanga, G.G., Haario, H.: Optimal control of two age structured malaria model with model parameter uncertainty. In: 11th World Congress on Computational Mechanics (2013)
16. Lashari, A.A., Aly, S., Hattaf, K., Zaman, G., Jung, I.H., Li, X.Z.: Presentation of malaria epidemics using multiple optimal controls. *J. Appl. Math.* (2012)
17. Otieno, G., Koske, J.K., Mutiso, J.M.: Transmission dynamics and optimal control of malaria in Kenya. *Hindawi Discret. Dyn. Nat. Soc.* **2016**, 1–27 (2016)
18. Athithan, S., Ghosh, M.: Stability analysis and optimal control of a malaria model with larvivorous fish as biological control agent. *Appl. Math. Inf. Sci.* **9**(4), 1893–1913 (2015)
19. Silva, C.J., Torres, D.F.M.: An optimal control approach to malaria prevention via insecticide-treated nets. Hindawi Publishing Corporation, Conference Papers in Mathematics, Volume 2013 (2013)
20. Silva, C.J., Torres, D.F.M., Venturino, E.: Optimal spraying in biological control of pests. *Math. Model. Nat. Phenom.* **12**(3), 51–64 (2017)
21. Tchuente, J.M., Khamis, S.A., Augusto, F.B., Mpeshe, S.C.: Optimal control and sensitivity analysis of an influenza model with treatment and vaccination. *Acta Biotheor.* **59**, 1–28 (2011)
22. Mwamtobe, P.M.M.: Optimal (Control of) intervention strategies for malaria epidemic in Karonga District, Malawi. Ph.D. thesis (2014)
23. Augusto, F.B., Elmojtaba, I.M.: Optimal control and cost-effective analysis of malaria/visceral leishmaniasis co-infection. *PLoS ONE* **12**(2), 1–31 (2017)
24. White, M.T., Conteh, L., Cibulskis, R., Ghani, A.C.: Costs and cost-effectiveness of malaria control interventions - a systematic review. *Malaria J.* **10**(1), 1–14 (2011)
25. Pitt, C., Ndiaye, M., Conteh, L., Sy, O., Ba, E., Cisse, B., Gomis, J.F., Gaye, O., Ndiaye, J., Milligan, P.J.: Large-scale delivery of seasonal malaria chemoprevention to children under 10 in Senegal: an economic analysis. *Health Policy Plann.* **32**(9), 1256–1266 (2017)

Effect of Genetic Defects in a Cortical Circuit Model Associated with Childhood Absence Epilepsy



Maliha Ahmed and Sue Ann Campbell

Abstract Childhood absence epilepsy is a pediatric epilepsy disorder associated with mutations in genes which encode ion channels including sodium channels. The thalamocortical circuit is considered to play an important role in the pathophysiology of absence seizures, exhibiting the ability to generate oscillations of different frequencies. The purpose of our investigation was to explore some of the genetic mutations that alter the function of individual neurons in the cerebral cortex, giving rise to an epileptic network. In particular, we investigated the consequence of these alterations on neuronal network activity associated with this disorder. In this regard, we created a small network consisting of deep layer cortical pyramidal neurons and an interneuron, each described by a single-compartment Hodgkin-Huxley style model. We investigated factors that convert a normal network into a hyperexcitable one, including impairment of $GABA_A$ synapses and sodium channel defects resulting from mutations in genes encoding sodium channels. Our model agrees with experimental results indicating the role of GABA impairment in generating a hyperexcitable network. Our results also suggest that the co-existence of multiple sodium channel mutations alters individual neuronal function to increase or decrease the likelihood of the network exhibiting seizure-like behaviour.

Keywords Childhood absence epilepsy · Computational model · Thalamocortical network · Sodium channel defects · Hodgkin-huxley · Genetic mutations

1 Introduction

Childhood absence epilepsy (CAE) is a common idiopathic pediatric epilepsy syndrome accounting for between 2 and 10% of all cases of epilepsy in children [1].

M. Ahmed (✉) · S. A. Campbell
Department of Applied Mathematics, University of Waterloo, Waterloo, Canada
e-mail: m243ahme@uwaterloo.ca

S. A. Campbell
e-mail: sacampbell@uwaterloo.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_44

It is characterized by brief episodes of impaired consciousness lasting about 10–15 s, and may occur hundreds of times a day. During these episodes, a child may stare blankly accompanied by an upward roll of their eyeballs, without any convulsive motor activity [2]. Although in most cases absence seizures resolve in adolescence, in about 20% of cases, children with CAE continue to have the condition which may progress in severity [3].

Clinically, the most common tool used for detection and classification of epilepsy type is an electroencephalogram (EEG). The procedure consists of tiny electrodes being placed on the patient's scalp which detect electrical charges resulting from brain cell activity. Thus, an EEG is capable of detecting abnormal patterns of activity. A 2.5–4 Hz spike-and-wave discharge (SWD) pattern, for example, is a signature of absence seizures in humans [3]. The spike-and-wave patterns refers to brief spikes (very fast oscillations) followed by a slower variation, called a slow wave [4]. Moreover, EEG patterns can be used an indicator of brain activity on a network level to study circuits involved in creating those patterns. Given the young age of patients and potentially invasive nature of study, genetic models of rats and mice (such as the GAERS and WAG/Rij rats) are commonly used to study absence epilepsy [5, 6]. The SWD frequency corresponding to absence seizures is higher in genetic rodent models, in the range of 7–11 Hz [5]. In addition, mathematical and computational modelling are powerful tools as well to understand the dynamics of brain activity during an epileptic seizure.

The thalamocortical circuit is an important component in the pathophysiology of absence seizures. This circuit consists of pathways connecting the thalamus and cerebral cortex, forming feedback loops between the two structures. The cortex, in particular, is integral in the development of spike-and-wave oscillations, and hence the focus in our work [7]. The human cerebral cortex is divided into six distinct layers. Cortical pyramidal neurons are the most numerous type of cortical neurons (mostly populated in layers 5 and 6) and are the main source of output of information from the cerebral cortex. Input from the thalamus mainly arrives in layer 4 and gets projected down to layers 5 and 6. It is then integrated and directed to appropriate cortical regions, as well as forming a feedback loop back to the thalamus [8].

In this paper, we use a small network of cortical neurons (illustrated in Fig. 1), each described by a single-compartment Hodgkin-Huxley style model, to explore factors that alter the function of individual neurons and give rise to an epileptic network associated with childhood absence epilepsy. In particular, we consider the following two factors: impairment of GABA synapses, and alterations of ion channels to simulate genetic mutations associated with CAE.

2 Model

All of the individual neuron models for this work were based on the thalamocortical network model by Traub et al. [9]. Our focus on the cortical component in the absence circuitry was motivated by rat models in which seizure initiation was found to be

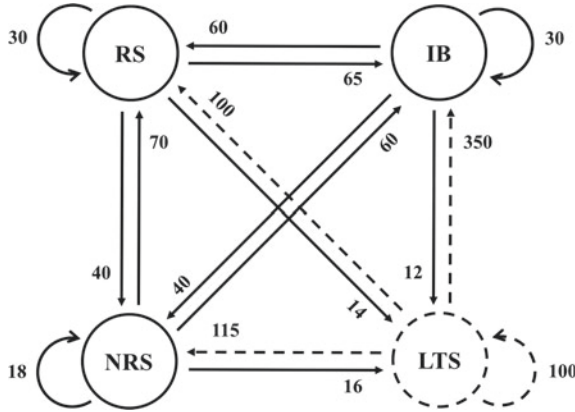


Fig. 1 Schematic of our cortical network model with excitatory (solid lines) and inhibitory (dashed lines) synapses. The numbers represent maximal synaptic conductance (in nS) for each synapse. Note: *RS*: layer 5 tufted regular spiking, *IB*: layer 5 tufted intrinsically bursting, *NRS*: layer 6 non-tufted regular spiking (pyramidal neurons), *LTS*: deep layer low-threshold spiking (interneuron)

associated with neurons in deep layers of the cortex [10]. Thus, we only included the layer 5 (tufted intrinsically bursting (IB) and regular spiking (RS)) and layer 6 (non-tufted regular spiking (NRS)) pyramidal neurons and the deep layer low-threshold spiking (LTS) interneuron in our model.

The following equation is the principal equation for each neuron describing the variation in time of membrane potential, $V(t)$:

$$C_m \frac{dV}{dt} = I_{hold} + I_{app} - I_h - I_{Naf} - I_{Nap} - I_{Kdr} - I_{Ka} - I_{K2} - I_{Km} - I_{Kc} - I_{Kahp} - I_{CaT} - I_{CaL} - I_{Leak} - I_{syn} \tag{1}$$

where C_m is the membrane capacitance (in $\mu F/cm^2$) with a value of 0.9 for all pyramidal neurons and 1.0 for the deep LTS interneuron [9]. I_{hold} is an input current used to set the resting membrane potential, I_{app} is a constant applied current, and I_{syn} is a sum of AMPA- and GABA-mediated synaptic currents. A consistent set of units were maintained such that voltage is given in mV, current in $\mu A/cm^2$, membrane conductance densities in mS/cm^2 , and time in msec. The incorporated Ca^{2+} , Na^+ , and K^+ currents included in the model are summarized in Table 1.

All equations for the currents were modelled in a Hodgkin-Huxley formalism (given in Table 1), where m and h are gating variables for ion channels. Ion channels are typically regulated by voltage-dependent gates which can go between a closed and open state. Let \mathbf{p} represent the fraction of open gates (e.g. m or h), then,

$$\frac{d\mathbf{p}}{dt} = \frac{\mathbf{p}_\infty(V) - \mathbf{p}}{\tau(V)} \tag{2}$$

Table 1 Description of currents used in our neuron models

Current	Current description	Current equation
I_h	Anomalous rectifier/ hyperpolarization-activated mixed cation current	$g_h \cdot (V - V_h) \cdot m_h$
I_{Naf}	Fast and transient inactivating Na^+ current	$g_{Naf} \cdot (V - V_{Na}) \cdot m_{Naf}^3 \cdot h_{Naf}$
I_{Nap}	Persistent Na^+ current	$g_{Nap} \cdot (V - V_{Na}) \cdot m_{Nap}$
I_{Kdr}	Delayed rectifier K^+ current	$g_{Kdr} \cdot (V - V_K) \cdot m_{Kdr}^4$
I_{Ka}	Transient inactivating K^+ current	$g_{Ka} \cdot (V - V_K) \cdot m_{Ka}^4 \cdot h_{Ka}$
I_{K2}	Slowly activating and inactivating K^+ current	$g_{K2} \cdot (V - V_K) \cdot m_{K2} \cdot h_{K2}$
I_{Km}	Muscarinic receptor-supressed K^+ current	$g_{Km} \cdot (V - V_K) \cdot m_{Km}$
I_{Kc}	Fast voltage and Ca^{2+} -dependent K^+ current	$g_{Kc} \cdot (V - V_K) \cdot m_{Kc} \cdot$ $\min(0.004 \cdot \chi, 1.0)$
I_{Kahp}	Slow Ca^{2+} -dependent K^+ current	$g_{Kahp} \cdot (V - V_K) \cdot m_{Kahp}$
I_{CaT}	Low-threshold inactivating Ca^{2+} current	$g_{CaT} \cdot (V - V_{Ca}) \cdot m_{CaT}^2 \cdot h_{CaT}$
I_{CaL}	High-threshold Ca^{2+} current	$g_{CaL} \cdot (V - V_{Ca}) \cdot m_{CaL}^2$
I_{Leak}	Leak current	$g_{Leak} \cdot (V - V_{Leak})$

where

$$\mathbf{p}_\infty(V) = \frac{1}{1 + \exp\left(\frac{-V + V_{1/2}}{k}\right)} \quad (3)$$

Note that $\tau(V)$ is the time constant, $\mathbf{p}_\infty(V)$ is the steady state value, $V_{1/2}$ half-activation voltage and k is the slope of voltage dependence [11]. Equations describing each gating variable as well as the details for calcium (denoted by χ) dynamics in our model are from [9], and a summary can be found in Sect. 5.2 of [12]. The maximal conductance densities for each current are given in Table 6.1 in [12].

Like ionic currents, synaptic currents can be modelled using a gating variable $s(t)$ which denotes the fraction of open synaptic channels at time t . According to the synaptic model developed by Destexhe et al. in 1994 [13], the synaptic current is described by

$$I_{syn} = \bar{g}_{syn} \cdot s(t) \cdot (V_{post} - V_{rev}) \quad (4)$$

where \bar{g}_{syn} is the maximal synaptic conductance, V_{post} is the membrane potential of the postsynaptic neuron, and V_{rev} is the reversal potential of the synapse. For AMPA and GABA synapses, V_{rev} was set to 0 mV and -75 mV, respectively [11]. Moreover, the fraction of bound receptors, $s(t)$ satisfies,

$$\frac{ds}{dt} = \alpha[T](1 - s) - \beta s \quad (5)$$

where $[T]$ represents the neurotransmitter concentration, and α and β are the forward and backward rate constants describing neurotransmitter binding. For AMPA-mediated synapses such as pyramidal (PYR)–PYR and PYR–interneuron, we set $\alpha = 1.4493$ and $2.8985 \text{ mM}^{-1}\text{ms}^{-1}$ respectively, and $\beta = 0.2173$ and 0.4346 ms^{-1} respectively. For GABA-mediated synapses, we set $\alpha = 5 \text{ mM}^{-1}\text{ms}^{-1}$ and $\beta = 0.125 \text{ ms}^{-1}$ [11, 14, 15]. Since neurotransmitter release is dependent on the presynaptic voltage, it is assumed to take the following form [11]:

$$[T](V_{pre}) = \frac{T_{max}}{1 + \exp\left(\frac{-(V_{pre} - V_T)}{K_p}\right)} \quad (6)$$

where T_{max} is the maximal concentration of the neurotransmitter in the synaptic cleft, V_{pre} is the presynaptic voltage, K_p is the steepness of voltage dependence, and V_T is the voltage at which the function is half-activated. We use the values suggested by Destexhe et al. namely, $T_{max} = 1 \text{ mM}$, $V_T = 2 \text{ mV}$ and $K_p = 5 \text{ mV}$ [13]. For connectivity of all neurons in our model, the maximal synaptic conductances were set for each synapse as given in Fig. 1.

3 Isolated Single Neurons

We began our investigation by first creating a definition of normal neuronal network activity in the cortex. In this regard, we first determined the firing behaviour of individual neurons. Thus, we compared the firing rate versus input current (f-I curves) for all neuron types, as given in Fig. 2a. Comparing the firing curves for layer 5 RS and layer 6 RS neurons, it can be seen that both neurons require similar amounts of stimulating current to initiate firing. On the contrary, the layer 5 IB neuron required the largest input current to initiate firing, while the interneuron required the smallest current to both initiate firing, and produced more action potentials for most inputs. The firing behaviour of all neurons was noted to be comparable to firing patterns reported in experimental works [16]. As isolated single neurons, simulations of each neuron with constant current input is given in Fig. 2b.

Next, with all synapses active, we defined a default state of our network. Network behaviour with input current to all pyramidal neurons and none to the interneuron, consists of synchronized regular firing of all neurons as seen in Fig. 2c. It is interesting to note the nature of firing (regular spiking) of all neurons given that one of the neurons has bursting properties by default. We attribute this result to the negative feedback from the inhibitory neuron, and its ability to suppress bursts into single spikes. Although not shown here, when the interneuron has a nonzero applied current, firing of all pyramidal neurons is fully suppressed while the interneuron spikes

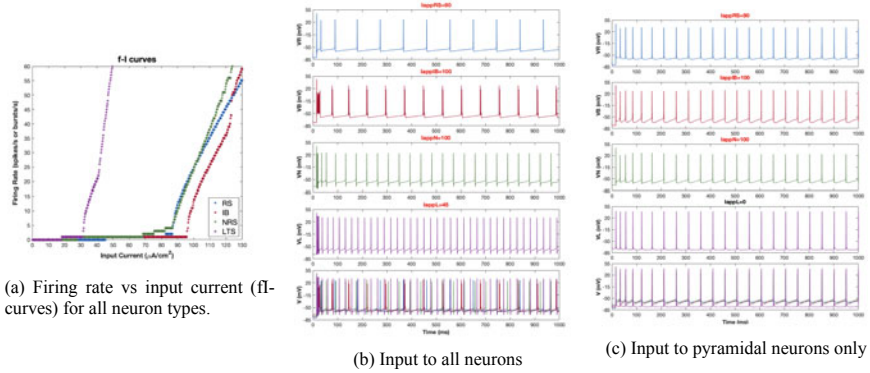


Fig. 2 Network with no connectivity between neurons for (a)–(b) and fully-connected network for (c). Units of applied current are $\mu A/cm^2$

repetitively. The reason for this is that the pyramidal neurons are unable to recover from the hyperpolarizing response before the interneuron fires again.

4 Modeling Impairment of GABA Synapses

Investigating factors that convert a normal network into a hyperexcitable one, we considered impairment of $GABA_A$ synapses, following evidence of impaired GABA inhibition in the case of absence epilepsy. Furthermore, it has been reported that EEG recordings from the cortex following blockade of $GABA_A$ receptors with penicillin exhibit synchronous spike-and-wave discharges (SWDs), indicating the crucial role of disinhibition in generation of SWDs. Thus, it was a reasonable approach for us to model disinhibition by reducing the strength of the maximal GABA conductance for synapses from the interneuron to all pyramidal neurons. Figure 3 shows the distribution of inter-spike intervals as inhibition is progressively reduced from 50 (the distribution is nearly the same from 0 to 45%) to 100%. It can be noted that the distribution of all neurons transforms from being uni-modal to bi-modal. This is indicative of the appearance of burst firing of all neurons as the modes corresponding to short and long intervals denote inter-spike and inter-burst intervals, respectively. The appearance of bursts as disinhibition is increased is confirmed by inspecting time series of individual neurons (not shown here).

In the presence of inhibition, there exist strong inhibitory feedback loops from the interneuron to all pyramidal neurons. Consequently, external stimuli to pyramidal neurons together with excitatory synapses with other pyramidal neurons ensures strong mutual excitation, including excitation to the interneuron. As a result, pyramidal neuron firing is rapidly dampened by strong inhibitory effects of the interneuron, making it difficult for the pyramidal neurons to produce bursts. However, in the

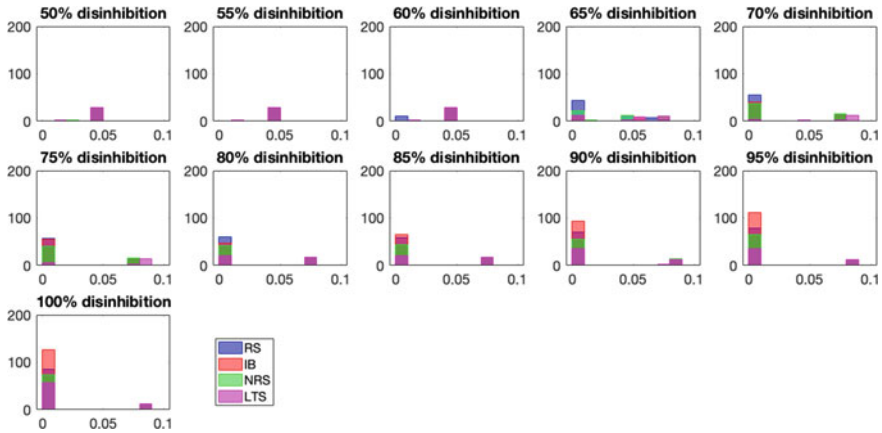


Fig. 3 Distribution of inter-spike intervals of firing with varying strengths of disinhibition

absence of GABAergic inhibition, excitatory feedback from all neurons is capable of depolarizing the neurons over their threshold potentials for a period of time that is sufficient to produce bursts. In particular, analogous to appearance of SWDs in a thalamocortical network, higher-frequency tonic spiking (~ 24 Hz) in our cortical network switch to slow bursting (~ 10 Hz).

An analysis of all gating variables for individual neurons revealed the role of the slow activation of calcium-dependent potassium currents such as I_{K_m} , I_{K_2} in termination of bursts (details can be found in [12]), and thus, provided insight on possible mechanisms underlying burst discharges.

5 Modeling Sodium Channel Mutations

Exploring hyperexcitability of our cortical network further, we modelled defects in sodium channel function based on literature on *SCN2a* and *SCN8a* mutations associated with absence epilepsy. These genes encode the Nav1.2 and Nav1.6 sodium channels, respectively. These channels are among the most prominent sodium channels present in the axon and dendrites of cortical neurons, including layer 5 pyramidal neurons [17]. Although there exist mutations in genes encoding h-channels and calcium channels, we limited our attention to sodium channel defects as voltage-gated sodium channels are one of the basic constituents of an action potential.

Based on literature, we introduced three different variants of *SCN8a* in the inhibitory interneuron, while the *SCN2a* variant was only introduced in all pyramidal neurons. In particular, the effects of mutations on these channel types were such that, alterations to the interneuron were specific to the fast Na^+ current, I_{Na_f} and alterations to pyramidal neurons were specific to the persistent Na^+ current, I_{Na_p} .

In 2018, it was discovered by Ogiwara et al. [18] that mice with a knockout mutation, $SCN2a^{KO/+}$ showed electrocorticography (ECoG) and electromyography (EMG) recordings resembling absence seizures. In their study, it was discovered that Nav1.2 deficiency, particularly in excitatory neurons, was causing the epileptic phenotype. Accordingly, we modelled this mutation by reducing the maximal conductance of the persistent sodium current in pyramidal neurons by 40% [18].

In 2009, a study by Papale et al. [19] reported mutations in the $SCN8a$ gene with effects associated with absence seizures in a mouse model. Membranes containing Nav1.6 channels are more excitable than membranes containing other sodium channels, and thus, impairment in Nav1.6 function is known to result in reduced neural firing. In their study, the effect of three different $SCN8a$ mutants was studied, namely $SCN8a^{V929F}$, $SCN8a^{med}$ and $SCN8a^{med-jo}$. The *med* variant was a null mutation which we modelled by setting the conductance of I_{Na_f} current to 0. The other two variants, *med-jo* and $V929F$ resulted in alteration in voltage dependence of activation and inactivation. We modelled these by introducing absolute depolarizing and hyperpolarizing shifts in voltage-dependence of activation and inactivation, respectively (values given in Table 2). We chose to model alterations of Nav1.6 channels in the interneuron since there is evidence of Nav1.6 expression in interneurons in the cortex, and a decrease in activity of inhibitory neurons could result in a net effect of excitation and hypersynchrony in the cortex (as is the case in the thalamus [20]) associated with absence seizures.

The effect induced by $SCN2a$ mutations in individual pyramidal neurons, was such that there was almost no change in the peak amplitudes, and the firing frequency was reduced for all neurons. However, the effect of all three variants of $SCN8a$ on the interneuron (in isolation) was such that firing was completely abolished. Furthermore, effects of $SCN2a$ and $SCN8a$ variants on the full network model were tested individually as well as in combination with each other. Introduction of $SCN2a$ variant only showed minimal changes in network behaviour as all neurons continued to exhibit synchronized regular spiking (not shown here). However, as shown in Fig. 4, introduction of $SCN8a^{med}$ and $SCN8a^{med-jo}$ individually as well as in combination with $SCN2a^{KO/+}$ resulted in burst discharges of all pyramidal neurons. The effect of $SCN8a^{V929F}$ was to produce alternating bursts and single spikes, and in combination with $SCN2a^{KO/+}$, the weakened feedback from pyramidal neurons to the interneuron produced a net effect of single spikes.

Table 2 Absolute shifts in half-activation voltages for variant Nav1.6 channels

Channel type	Voltage-dependence of activation		Voltage-dependence of inactivation	
	$\Delta V_{1/2}$	Δk	$\Delta V_{1/2}$	Δk
$Nav1.6^{med-jo}$	14.1	0.3	6.8	-0.7
$Nav1.6^{V929F}$	5.87	1.84	-11.79	-1.21

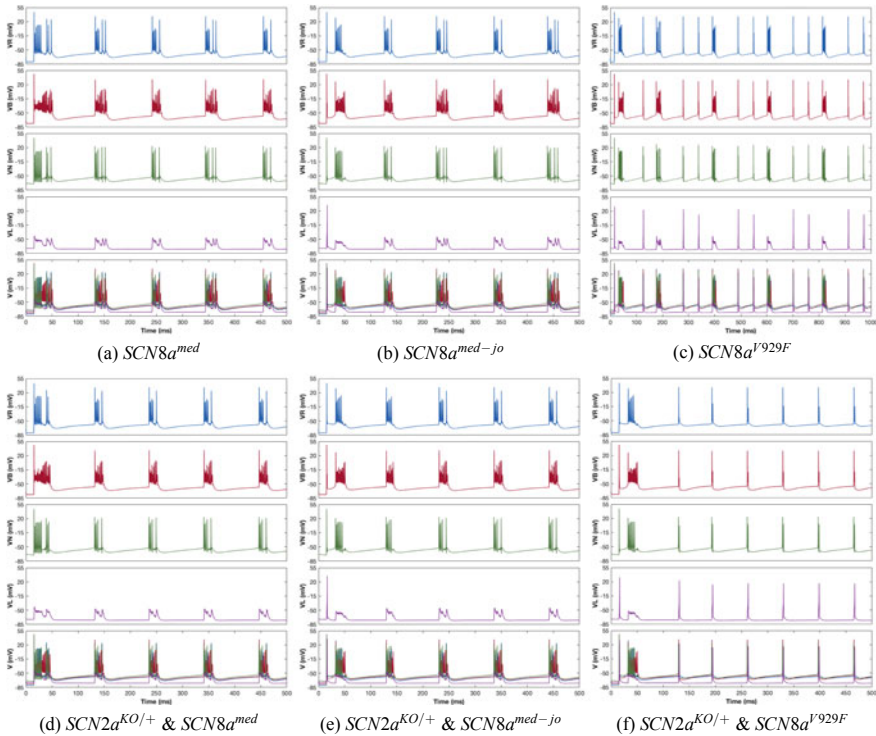


Fig. 4 Network behaviour following the introduction of *SCN2a* and *SCN8a* mutant variants. **a–c** introduction of *SCN8a* mutation only. **d–f** both *SCN2a* and *SCN8a* mutant variants included

6 Conclusion and Future Directions

Childhood absence epilepsy is suspected of resulting from mutations in genes which encode ion-channels, including particular sodium channels. Our purpose in this work was to model the effect of some of the mutations documented in the literature and explore how alterations in individual cortical neurons contributes to network activity associated with CAE. A cortical network consisting of pyramidal neurons and an interneuron was used to define normal and abnormal cortical network activity. We modelled impairment of GABA synapses, following evidence of this for absence epilepsy. The result was hyperexcitability of our cortical network, and in particular, the appearance of network-induced burst discharges, analogous to the appearance of SWDs in a thalamocortical network. By examining gating variables for all neurons in the “normal” and “abnormal” states, we showed that bursts were sustained by an influx of calcium and terminated by slow calcium-dependent potassium currents, such as I_{Km} and I_{K2} , consistent with other models [21].

Based on literature on *SCN2a* and *SCN8a* mutations associated with absence epilepsy, we modelled sodium channel defects, particularly for Nav1.2 and Nav1.6

sodium channels. To our knowledge, modelling *SCN8a* variants in cortical interneurons has not been done previously, and in cortical networks, it was an important aspect to study. The introduction of three variants (*med*, *med - jo* and *V929F*) into the network model transformed the network into a hyperexcitable (epileptic) state. While *med* and *med - jo* gave similar effects on the network, the effect of *V929F* was relatively weaker. An analysis of gating variables as in the case of impaired GABA synapses revealed a similar underlying mechanism producing burst discharges of all neurons. From this result, we can conclude that the functional consequence of *SCN8a* mutations affecting interneuron function and thus, impacting neuronal firing upstream in the absence circuitry, is similar to the effect of impaired GABA function. We also modelled the effect of *SCN2a* knockout mutation on cortical pyramidal neurons. Although this defect had very little effect on the *med* and *med - jo* cases, it almost normalized *V929F* activity. Contrary to previous experimental results [18], we did not observe any hyperexcitability of our cortical network with a *SCN2a* defect. One possible explanation for this is the absence of connectivity with the thalamus. While *SCN2a* caused decreased excitability of cortical pyramidal neurons, in a larger network context this could cause decreased input to thalamic interneurons, reducing inhibition to thalamic excitatory neurons leading to a net effect of hyperexcitability [18].

Overall, these results were able to provide insight on the role of genetic mutations in absence epilepsy and possible underlying mechanisms giving rise to hyperexcitable cortical networks. Our results also suggest the importance of considering thalamic connectivity in future works to obtain a full picture, especially in exploring genetic manipulations in cortical circuits and their functional effect on the thalamo-cortical network.

References

1. Posner, E.: Absence seizures in children. *BMJ Clin. Evid.* **0317** (2008)
2. Buchhalter, J.: Treatment of childhood absence epilepsy: an evidence-based answer at last!. *Epilepsy Curr.* **11**(1), 12–15 (2011)
3. Crunelli, V., Leresche, N.: Childhood absence epilepsy: genes, channels, neurons and networks. *Nat. Rev. Neurosci.* **3**(5), 371–382 (2002)
4. Epilepsy Society. A closer look at EEG. <https://www.epilepsysociety.org.uk/closer-look-eeeg/#W7UImxNKiog>
5. Meeren, H.K.M., Pijn, J.P.M., Van Luijtelaar, E.L.J.M., Coenen, A.M.L., Lopes da Silva, F.H.: Cortical focus drives widespread corticothalamic networks during spontaneous absence seizures in rats. *J. Neurosci.* **22**(4), 1480–1495 (2002)
6. Polack, P., Guillemain, I., Hu, E., Deransart, C., Depaulis, A., Charpier, S.: Deep layer somatosensory cortical neurons initiate spike-and-wave discharges in a genetic model of absence seizures. *J. Neurosci.* **27**(24), 6590–6599 (2007)
7. Danober, L., Deransart, C., Depaulis, A., Vergnes, M., Marescaux, C.: Pathophysiological mechanisms of genetic absence epilepsy in the rat. *Prog. Neurobiol.* **55**(1), 27–57 (1998)
8. Constantinople, C.M., Bruno, R.M.: Deep cortical layers are activated directly by thalamus. *Science* **340**(6140), 1591–1594 (2013)

9. Traub, R.D., Contreras, D., Cunningham, M.O., Murray, H., LeBeau, F.E., Roopun, A., Whittington, M.A., et al.: Single-column thalamocortical network model exhibiting gamma oscillations, sleep spindles, and epileptogenic bursts. *J. Neurophys.* **93**, 2194–2232 (2005)
10. Depaulis, A., Charpier, S.: Pathophysiology of absence epilepsy: insights from genetic models. *Neurosc. Lett.* **667**, 53–65 (2018)
11. Ermentrout, B.G., Terman, D.H., Antman, S.S., Marsden, J.E., Sirovich, L., Wiggins, S.: *Mathematical Foundations of Neuroscience*. Springer, New York (2010)
12. Ahmed, M.: Model for a cortical circuit associated with childhood absence epilepsy. MMath thesis. University of Waterloo (2019)
13. Destexhe, A., Mainen, Z.F., Sejnowski, T.J.: Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism. *J. Comput. Neurosci.* **1**(3), 195–230 (1994)
14. Salin, P.A., Prince, D.A.: Spontaneous GABA-A receptor-mediated inhibitory currents in adult rat somatosensory cortex. *J. Neurophys.* **75**(4), 1573–1588 (1996)
15. Hestrin, S.: Different glutamate receptor channels mediate fast excitatory synaptic currents in inhibitory and excitatory cortical neurons. *Neuron* **11**(6), 1083–1091 (1993)
16. Williams, S.R., Stuart, G.J.: Mechanisms and consequences of action potential burst firing in rat neocortical pyramidal neurons. *J. Physiol.* **521**, 467–482 (1999)
17. Child, N.D., Benarroch, E.E.: Differential distribution of voltage-gated ion channels in cortical neurons. *Neurology* **82**(11), 989–999 (2014)
18. Ogiwara, I., Miyamoto, H., Tatsukawa, T., Yamagata, T., Nakayama, T., Atapour, N., Yamakawa, K., et al.: Nav1.2 haploinsufficiency in excitatory neurons causes absence-like seizures in mice. *Commun. Biol.* **1**(96) (2018)
19. Papale, L.A., Beyer, B., Jones, J.M., Sharkey, L.M., Tufik, S., Epstein, M., Escayg, A., et al.: Heterozygous mutations of the voltage-gated sodium channel SCN8a are associated with spike-wave discharges and absence epilepsy in mice. *Human Mol. Genet.* **18**(9), 1633–1641 (2009)
20. Makinson, C.D., Tanaka, B.S., Sorokin, J.M., Wong, J.C., Christian, C.A., Goldin, A.L., Huguenard, J.R.: Regulation of thalamic and cortical network synchrony by scn8a. *Neuron* **93**(5), 1165–1179 (2017)
21. Destexhe, A., Sejnowski, T.J.: The initiation of bursts in thalamic neurons and the cortical control of thalamic sensitivity. *Philos. Trans. R. Soc. Lond. Series B Biol. Sci.* **357**, 1649–1657 (2002)

Operator Splitting for the Simulation of Aqueous Humor Thermo-Fluid-Dynamics in the Anterior Chamber



Farah Abdelhafid, Giovanna Guidoboni, Naoki Okumura, Noriko Koizumi, and Sangly P. Srinivas

Abstract This work presents a numerical scheme based on operator splitting for the thermo-fluid-dynamical simulation of aqueous humor flow in the anterior chamber of the human eye. The stability properties of the scheme are investigated theoretically. Numerical results are presented for different postures and different external temperatures.

Keywords Operator Splitting · Aqueous humor dynamics · Thermo-fluid dynamics

1 Introduction

This work presents a numerical method based on operator splitting for the thermo-fluid-dynamical simulation of aqueous humor flow in the anterior chamber of the human eye. Aqueous humor flow is very important for several physiological functions, including establishing intraocular pressure and provide nutrients and oxygen

F. Abdelhafid

Department of Mathematics, University of Missouri, Columbia, MO, USA

e-mail: fyatzd@mail.missouri.edu

G. Guidoboni (✉)

Department of Electrical Engineering and Computer Science, Department of Mathematics, University of Missouri, Columbia, MO, USA

e-mail: guidobonig@missouri.edu

N. Okumura · N. Koizumi

Department of Biomedical Engineering, Doshisha University, Kyotanabe, Japan

e-mail: nokumura@mail.doshisha.ac.jp

N. Koizumi

e-mail: nkoizumi@mail.doshisha.ac.jp

S. P. Srinivas

Indiana University Bloomington, Bloomington, IN, USA

e-mail: srinivas@indiana.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_45

to the avascular structures in the eye, such as the cornea. The features of aqueous humor flow are known to be influenced by several factors, including posture and outside temperature [4]. Gaining knowledge of the quantitative effect of changes in these factors on aqueous humor flow will enable the utilization of such factors to optimize/design therapeutical interventions.

The aqueous humor, whose volume is approximately 300 μL , is secreted at a rate of approximately 2.5 $\mu\text{L}/\text{min}$ into the posterior chamber by the ciliary epithelium, escapes into the anterior chamber through the pupil and exits the eye via the trabecular meshwork in a segmental fashion around the limbus. It also undergoes mixing in the anterior chamber because of convection currents induced by the temperature difference between the corneal surface, which is exposed to the temperature of the external ambient, and the vascularized iris, which is at body temperature. The flow field of the aqueous humor affects (a) the residence time and distribution kinetics of topical drugs, (b) optimal placement of drug delivery implants in the anterior chamber [10], and (c) settlement of the inflammatory cells on the endothelial surface (e.g., Krukenberg's spindle) [1]. Flow analysis can also aid us in the development of technologies to force rapid sedimentation of corneal endothelial cells following cell injection therapy being promoted for the treatment of bullous keratopathy [11]. Thermo-fluid-dynamical studies of aqueous humor flow have been considered by other authors, see for example [1, 4, 9, 13]. However, in most cases the simulations were conducted by means of commercial software, which often provides user-friendly interfaces with limited access to the numerical strategy for the solution of coupled problems. Operator splitting has proved to be very effective for the numerical solution of time-dependent multi-physics problems arising in fluid-dynamics [6], including fluid-structure interactions [7] and multiscale 3d-0d coupling [2]. This paper explores the feasibility of such an approach to the thermo-fluid-dynamical study of aqueous humor flow as a first step towards more complex simulations including additional physical mechanisms.

2 Mathematical Model

We implemented the geometry proposed in [3] in two spatial dimensions via Gmsh [5], see Fig. 1. Let us denote by Ω the two-dimensional domain representing the anterior chamber of the human eye. The boundary of Ω , denoted by $\partial\Omega$, can be written as $\partial\Omega = \Sigma_{PC} \cup \Sigma_I \cup \Sigma_{TM} \cup \Sigma_C \cup \Sigma_P$, where the subscripts PC , TM , C , I and P represent the posterior chamber (indicated as *Inlet* in Fig. 1), the trabecular meshwork (indicated as *Outlet* in Fig. 1), the cornea, the iris and the pupil, respectively. In Fig. 1, the portions of the boundary are marked by the points P_i , with $i = 1, \dots, 10$, whose coordinates are listed in Table 1.

The aqueous humor is modeled as a Newtonian viscous fluid in the Boussinesq approximation [9]. Let $(0, T)$, with $T > 0$, be a time interval. Then, in $\Omega \times (0, T)$ we solve the following problem

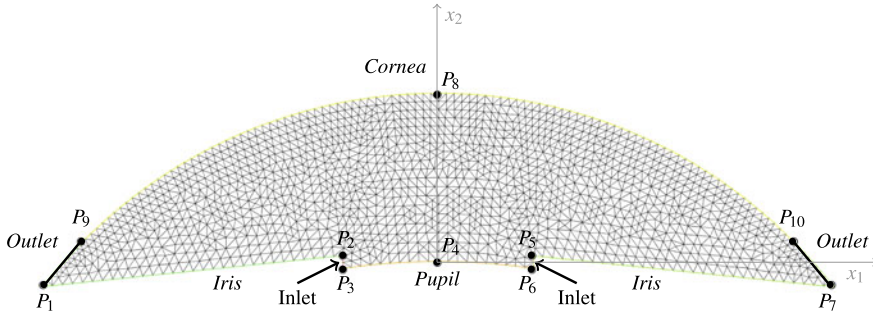


Fig. 1 Triangular mesh utilized for the space discretization of the domain. The schematic also indicates the different portions of the domain boundary

Table 1 Coordinates of the points marking the portions of the domain boundary

Point	x_1 (mm)	x_2 (mm)	Point	x_1 (mm)	x_2 (mm)
P_1	-6.2	-0.4	P_2	-1.5	0.087
P_3	-1.5	-0.113	P_4	0	0
P_5	1.5	0.087	P_6	1.5	-0.113
P_7	6.2	-0.4	P_8	0	2.62
P_9	-5.540	0.352	P_{10}	5.547	0.346

$$\nabla \cdot \underline{\mathbf{v}} = 0 \tag{1a}$$

$$\rho_0 C_p \left(\frac{\partial \theta}{\partial t} + \underline{\mathbf{v}} \cdot \nabla \theta \right) = \nabla \cdot (\kappa \nabla \theta) \tag{1b}$$

$$\rho_0 \left(\frac{\partial \underline{\mathbf{v}}}{\partial t} + \underline{\mathbf{v}} \cdot \nabla \underline{\mathbf{v}} \right) = -\nabla p + \mu \Delta \underline{\mathbf{v}} + \rho_0 \left(1 - \alpha(\theta - \theta_0) \right) \underline{\mathbf{g}} \tag{1c}$$

where $\underline{\mathbf{v}}$, p and θ represent the fluid velocity, pressure and temperature, respectively. The parameters μ , C_p , α and κ represent the dynamic viscosity, the specific heat, the coefficient of thermal expansion, and thermal conductivity of the aqueous humor, respectively. In addition, ρ_0 represents the fluid density at the reference temperature θ_0 . The vector $\underline{\mathbf{g}}$ represents the gravitational acceleration vector. Equations (1a)–(1c) express the balance of mass, energy and linear momentum governing the flow [14]. Numerical simulations will be conducted in the case of three different postures of interest for clinical applications, namely standing, laying supine and laying prone. These cases are characterized by a different definition of the gravitational acceleration vector, namely $\underline{\mathbf{g}} = [g, 0]^T$ in the standing position, $\underline{\mathbf{g}} = [0, -g]^T$ in the supine position and $\underline{\mathbf{g}} = [0, g]^T$ in the prone position, where g is the constant of gravitational acceleration. Initial conditions are specified for velocity and temperature as $\underline{\mathbf{v}} = \tilde{\underline{\mathbf{v}}}$ and $\theta = \tilde{\theta}$ for $\underline{\mathbf{x}} \in \Omega$ and $t = 0$. No slip boundary conditions are assumed for the fluid velocity on the cornea, iris and pupil. Nonhomogeneous Dirichlet conditions

are prescribed on the posterior chamber and trabecular meshwork, which represent the inlet and outlet portions of the domain boundary, respectively. Thus, the boundary conditions for the fluid velocity read as follows:

$$\underline{v} = \underline{0} \quad \text{on} \quad \Sigma_I \cup \Sigma_C \cup \Sigma_P \times (0, T) \quad (1d)$$

$$\underline{v} = -v_{in} \underline{n} \quad \text{on} \quad \Sigma_{PC} \times (0, T) \quad (1e)$$

$$\underline{v} = v_{out} \underline{n} \quad \text{on} \quad \Sigma_{TM} \times (0, T) \quad (1f)$$

where v_{in} and v_{out} are given values. The fluid temperature is assumed to be equal to the body temperature on the iris, posterior chamber, pupil and trabecular meshwork, so that we can write:

$$\theta = \theta_{body} \quad \text{on} \quad \Sigma_I \cup \Sigma_{PC} \cup \Sigma_P \cup \Sigma_{TM} \times (0, T) \quad (1g)$$

On the cornea Σ_C , the temperature is not assigned a priori but it results from the heat balance at the interface with the external ambient, so that we can write:

$$-\kappa \frac{\partial \theta}{\partial \underline{n}} = h_{amb}(\theta - \theta_{amb}) + \sigma \varepsilon (\theta^4 - \theta_{amb}^4) + E \quad \text{on} \quad \Sigma_C \times (0, T) \quad (1h)$$

where θ_{amb} is the temperature of the external ambient, which is assumed to be given, and the parameters h_{amb} , σ , ε and E represent the ambient convection coefficient, the Stefan-Boltzmann constant, the emissivity of cornea and the evaporation rate, respectively [13].

3 Numerical Method

We will solve problem (1) by combining the operator splitting method for the time discretization and the finite element method for the spatial discretization. When considering a differential system of the form:

$$\frac{\partial \underline{w}}{\partial t} = \mathcal{A}(\underline{w}) \quad \text{with} \quad \underline{w}(t = 0) = \underline{w}_0 \quad (2)$$

where \mathcal{A} is an operator acting on the unknown variable $\underline{w} = \underline{w}(\underline{x}, t)$, the operator splitting technique can be applied when \mathcal{A} can be conveniently written as the sum of simpler operators, namely $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2 + \dots + \mathcal{A}_M$. A detailed description of the operator splitting technique can be found in [6]. This technique becomes particularly advantageous when the problem at hand is of hyperbolic/parabolic type. In this case, by splitting the hyperbolic and parabolic parts of the operator we can design

a numerical scheme that enjoys unconditional stability with respect to the choice of the time step, as shown in Theorem 1.

Let us discretize the time interval $(0, T)$ by means of a uniform discretization based on the time step τ such that $t^n = n \tau$ for $n \geq 0$. In addition, let us use the notation $\varphi^n = \varphi(t^n)$ for any function φ . Then, the operator splitting method is utilized to approximate the solution of problem (1) by means of an algorithm comprising 4 steps solved sequentially.

For $n \geq 0$, assuming that \underline{v}^n and θ^n are known, solve:

Step 1: Find $\underline{v}^{n+1/4}$ and $\theta^{n+1/4}$ such that

$$\rho_0 C_p \frac{\partial \theta}{\partial t} = \nabla \cdot (\kappa \nabla \theta) \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (3a)$$

$$\rho_0 \frac{\partial \underline{v}}{\partial t} = \underline{\mathbf{0}} \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (3b)$$

with the boundary conditions

$$\theta = \theta_{body} \quad \text{on} \quad \Sigma_I \cup \Sigma_{PC} \cup \Sigma_P \cup \Sigma_{TM} \times (t^n, t^{n+1}) \quad (3c)$$

$$-\kappa \frac{\partial \theta}{\partial \underline{n}} = h_{amb}(\theta - \theta_{amb}) + \sigma \varepsilon (\theta^4 - \theta_{amb}^4) + E \quad \text{on} \quad \Sigma_C \times (t^n, t^{n+1}) \quad (3d)$$

and the initial conditions

$$\underline{v}(t^n) = \underline{v}^n, \quad \theta(t^n) = \theta^n \quad \text{in } \Omega \quad (3e)$$

and then set $\underline{v}^{n+1/4} = \underline{v}(t^{n+1}) = \underline{v}^n$ and $\theta^{n+1/4} = \theta(t^{n+1})$.

Step 2: Find $\underline{v}^{n+2/4}$, $p^{n+2/4}$ and $\theta^{n+2/4}$ such that

$$\nabla \cdot \underline{v} = 0 \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (4a)$$

$$\rho_0 C_p \frac{\partial \theta}{\partial t} = 0 \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (4b)$$

$$\rho_0 \frac{\partial \underline{v}}{\partial t} = -\nabla p + \mu \Delta \underline{v} + \rho_0 \left(1 - \alpha(\theta - \theta_0)\right) \underline{\mathbf{g}} \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (4c)$$

with the boundary conditions

$$\underline{v} = \underline{\mathbf{0}} \quad \text{on} \quad \Sigma_I \cup \Sigma_C \cup \Sigma_P \times (t^n, t^{n+1}) \quad (4d)$$

$$\underline{v} = -v_{in} \underline{n} \quad \text{on} \quad \Sigma_{PC} \times (t^n, t^{n+1}) \quad (4e)$$

$$\underline{v} = v_{out} \underline{n} \quad \text{on} \quad \Sigma_{TM} \times (t^n, t^{n+1}) \quad (4f)$$

and the initial conditions

$$\underline{v}(t^n) = \underline{v}^{n+1/4}, \quad \theta(t^n) = \theta^{n+1/4} \quad \text{in } \Omega \quad (4g)$$

and then set $\underline{v}^{n+2/4} = \underline{v}(t^{n+1}) = \underline{v}^n$, $p^{n+2/4} = p(t^{n+1})$ and $\theta^{n+2/4} = \theta(t^{n+1}) = \theta^{n+1/4}$.

Step 3 : Find $\underline{v}^{n+3/4}$ and $\theta^{n+3/4}$ such that

$$\rho_0 C_p \left(\frac{\partial \theta}{\partial t} + \underline{v}^{n+2/4} \cdot \nabla \theta \right) = 0 \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (5a)$$

$$\rho_0 \frac{\partial \underline{v}}{\partial t} = \underline{0} \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (5b)$$

with the boundary conditions

$$\theta = \theta^{n+2/4} \quad \text{on } \Sigma_- \times (t^n, t^{n+1}) \quad (5c)$$

where $\Sigma_- = \{ \underline{x} \in \partial \Omega : \underline{v}^{n+2/4} \cdot \underline{n} \leq 0 \}$ and the initial conditions

$$\underline{\theta}(t^n) = \theta^{n+2/4} \quad \text{in } \Omega \quad (5d)$$

and then set $\underline{v}^{n+3/4} = \underline{v}(t^{n+1}) = \underline{v}^{n+2/4}$ and $\theta^{n+3/4} = \theta(t^{n+1})$.

Step 4 : Find $\underline{v}^{n+4/4}$ and $\theta^{n+4/4}$ such that

$$\rho_0 C_p \frac{\partial \theta}{\partial t} = 0 \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (6a)$$

$$\rho_0 \left(\frac{\partial \underline{v}}{\partial t} + \underline{v}^{n+3/4} \cdot \nabla \underline{v} \right) = \underline{0} \quad \text{in } \Omega \times (t^n, t^{n+1}) \quad (6b)$$

with the boundary conditions

$$\underline{v} = \underline{v}^{n+3/4} \quad \text{on } \Sigma_- \times (t^n, t^{n+1}) \quad (6c)$$

where $\Sigma_- = \{ \underline{x} \in \partial \Omega : \underline{v}^{n+3/4} \cdot \underline{n} \leq 0 \}$ and the initial conditions

$$\underline{v}(t^n) = \underline{v}^{n+3/4} \quad \text{in } \Omega \quad (6d)$$

and then set $\underline{v}^{n+4/4} = \underline{v}(t^{n+1})$ and $\theta^{n+4/4} = \theta(t^{n+1}) = \theta^{3/4}$.

Finally set $\underline{v}^{n+1} = \underline{v}^{n+4/4}$, $p^{n+1} = p^{n+2/4}$ and $\theta^{n+1} = \theta^{n+4/4}$, advance n and repeat.

Theorem 1 *In the case of homogeneous boundary data and in the absence of external forces, the splitting algorithm consisting of the four steps (3)–(6) is unconditionally stable with respect to the choice of the time step.*

Proof Let us begin by considering problem (3) characterizing Step 1. By multiplying Eq. (3a) by θ and Eq. (3b) by \underline{v} , integrating over Ω and utilizing the boundary conditions (3c) and (3d) we obtain $d\mathcal{E}_1/dt + \mathcal{D}_1 = \mathcal{F}_1$ for $t \in (t^n, t^{n+1})$, where

$$\begin{aligned}\mathcal{E}_1 &= \frac{1}{2} \int_{\Omega} \rho_0 C_p \theta^2 d\Omega + \frac{1}{2} \int_{\Omega} \rho_0 |\underline{v}|^2 d\Omega \\ \mathcal{D}_1 &= \int_{\Omega} \kappa |\nabla \theta|^2 d\Omega + \int_{\Sigma_C} (h_{amb} \theta^2 - \sigma \varepsilon \theta^5 - E \theta) d\Sigma \\ \mathcal{F}_1 &= \int_{\partial\Omega \setminus \Sigma_C} \theta_{body} (\kappa \nabla \theta) \cdot \underline{n} d\Sigma + \int_{\Sigma_C} (h_{amb} \theta_{amb} + \sigma \varepsilon \theta_{amb}^4) \theta d\Sigma\end{aligned}$$

Let us now consider problem (4) characterizing Step 2. By multiplying Eq. (4b) by θ and Eq. (4c) by \underline{v} , integrating over Ω and utilizing the divergence free condition (4a) and the boundary conditions (4d)–(4e) we obtain $d\mathcal{E}_2/dt + \mathcal{D}_2 = \mathcal{F}_2$ for $t \in (t^n, t^{n+1})$, where

$$\begin{aligned}\mathcal{E}_2 &= \frac{1}{2} \int_{\Omega} \rho_0 C_p \theta^2 d\Omega + \frac{1}{2} \int_{\Omega} \rho_0 |\underline{v}|^2 d\Omega \\ \mathcal{D}_2 &= \int_{\Omega} \mu |\nabla \underline{v}|^2 d\Omega \\ \mathcal{F}_2 &= \int_{\Omega} \rho_0 (1 - \alpha(\theta - \theta_0)) \underline{g} \cdot \underline{v} d\Omega - \int_{\Sigma_{PC}} v_{in} (\underline{\sigma} \underline{n}) \cdot \underline{n} d\Sigma + \int_{\Sigma_{TM}} v_{out} (\underline{\sigma} \underline{n}) \cdot \underline{n} d\Sigma\end{aligned}$$

with $\underline{\sigma} = -p\underline{I} + \mu \nabla \underline{u}$. Let us now consider problem (5) characterizing Step 3. Eq. (5a) by θ and Eq. (5b) by \underline{v} , integrating over Ω , utilizing the divergence free condition (4a) satisfied by $\underline{v}^{n+2/4}$ and the boundary conditions (5c) we obtain $d\mathcal{E}_3/dt = \mathcal{F}_3$ for $t \in (t^n, t^{n+1})$, where

$$\begin{aligned}\mathcal{E}_3 &= \frac{1}{2} \int_{\Omega} \rho_0 C_p \theta^2 d\Omega + \frac{1}{2} \int_{\Omega} \rho_0 |\underline{v}|^2 d\Omega \\ \mathcal{F}_3 &= - \int_{\Sigma_{PC}} \frac{(\theta^{n+2/4})^2}{2} v_{in} d\Sigma + \int_{\Sigma_{TM}} \frac{\theta^2}{2} v_{out} d\Sigma\end{aligned}$$

Let us now consider problem (6) characterizing Step 4. Eq. (6a) by θ and Eq. (6b) by \underline{v} , integrating over Ω , utilizing the divergence free condition (4a) satisfied by $\underline{v}^{n+3/4}$ and the boundary conditions (6c) we obtain $d\mathcal{E}_4/dt = \mathcal{F}_4$ for $t \in (t^n, t^{n+1})$, where

$$\begin{aligned}\mathcal{E}_4 &= \frac{1}{2} \int_{\Omega} \rho_0 C_p \theta^2 d\Omega + \frac{1}{2} \int_{\Omega} \rho_0 |\underline{v}|^2 d\Omega \\ \mathcal{F}_4 &= - \int_{\Sigma_{PC}} \frac{|\underline{v}^{n+3/4}|^2}{2} v_{in} d\Sigma + \int_{\Sigma_{TM}} \frac{|\underline{v}|^2}{2} v_{out} d\Sigma\end{aligned}$$

If boundary data are homogeneous and external forces are absent, then $\mathcal{F}_i = 0$ for $i = 1, \dots, 4$. Thus, the initial conditions for each step allow us to write

$$\mathcal{E}_4(t^{n+1}) = \mathcal{E}_4(t^n) = \mathcal{E}_3(t^{n+1}) = \mathcal{E}_3(t^n) = \mathcal{E}_2(t^{n+1}) \leq \mathcal{E}_2(t^n) = \mathcal{E}_1(t^{n+1}) \leq \mathcal{E}_1(t^n)$$

yielding $\mathcal{E}_4(t^{n+1}) \leq \mathcal{E}_1(t^n)$ for $n \geq 0$, from which unconditional stability follows.

We emphasize that the four steps solved sequentially in the splitting scheme are coupled via the initial conditions. Theorem 1 shows that this coupling strategy is essential to preserve the natural energy balance determined by the physics of the problem and yield numerical stability. Even though the situation of homogeneous boundary data seems purely academic, it allows us to verify that the proposed numerical method based on operator splitting preserves the essentially dissipative nature of the system without introducing spurious sources of energy into the system.

4 Simulation Results

The splitting algorithm (3)–(6) was used to simulate the flow of aqueous humor in the human eye under different conditions of clinical interest. The values of the physical parameters adopted in the simulations are listed in Table 2. The parabolic problem in Step 1 was discretized in time via a one step Backward Euler scheme and a finite element discretization utilizing P2 elements. The Stokes problem in Step 2 was discretized in time via a one step Backward Euler scheme and a finite element discretization utilizing P1-P2 elements with stabilization. The advective problems in Steps 2 and 3 were solved via the Characteristics-Galerkin Method implemented in the function `convect` built in FreeFem++ [8]. We emphasize that: (i) our current implementation is only first-order accurate in time, but it could be made second order with Strang symmetrization, and (ii) the advective steps could be solved via a conservative wave-like method [6].

Figures 2, 3 and 4 report the simulated temperature profiles and flow streamlines in the case of supine, prone and standing positions when the external temperature is held constant at $\theta_{amb} = 25\text{C}$. The simulations confirm that changes in posture have a nontrivial effect on temperature and velocity profiles. For example, the temperature distribution is very similar in the supine and prone positions, but the internal fluid vortices rotate in opposite directions. Interestingly, the two vortices that characterize the supine and prone positions are not present when standing. Changes in the temperature external to the cornea strongly influences the flow magnitude, as shown in Figs. 5, 6 and 7 for θ_{amb} equal to 45C and 15C in different postures.

Table 2 Summary of model parameters

Parameter	Symbol	Value	Unit
Reference temperature	θ_0	298	K
Fluid density at θ_0	ρ_0	10^{-3}	g mm^{-3}
Specific heat	C_p	4.2×10^9	$\text{Jg}^{-1} \text{K}^{-1}$
Thermal conductivity	κ	578×10^3	$\text{W mm}^{-1} \text{K}^{-1}$
Coefficient of linear expansion of aqueous humour	α	3×10^{-4}	K^{-1}
Dynamic viscosity	μ	1.08×10^3	$\text{g mm}^{-1} \text{s}^{-1}$
Ambient temperature	θ_{amb}	298	K
Ambient convection coefficient	h_{amb}	1×10^4	$\text{W mm}^{-2} \text{K}^{-1}$
Stefan-Boltzmann constant	σ	5.67×10^{-5}	$\text{W mm}^{-2} \text{K}^{-4}$
Emissivity of cornea	ε	0.975	
Evaporation rate	E	4×10^4	$\text{W m}^{-2} \text{K}^{-4}$

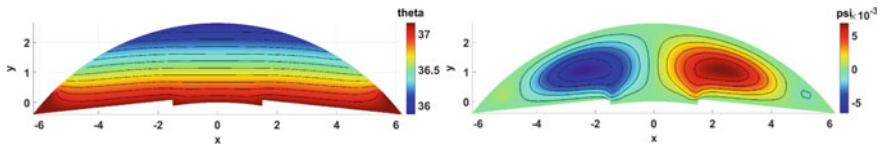


Fig. 2 *Supine position, $\theta_{amb} = 25\text{C}$. Temperature profile (Left) and flow streamlines (Right)*

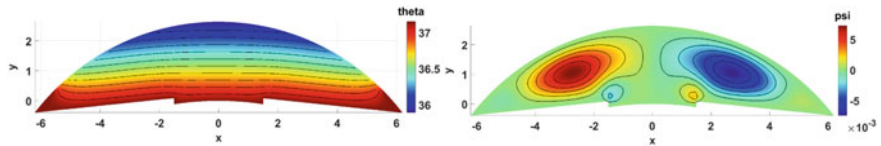


Fig. 3 *Prone position, $\theta_{amb} = 25\text{C}$. Temperature profile (Left) and flow streamlines (Right)*

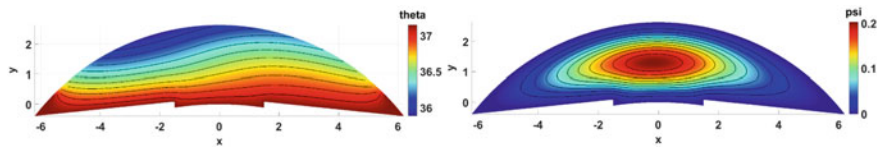


Fig. 4 *Standing position, $\theta_{amb} = 25\text{C}$. Temperature profile (Left) and flow streamlines (Right)*

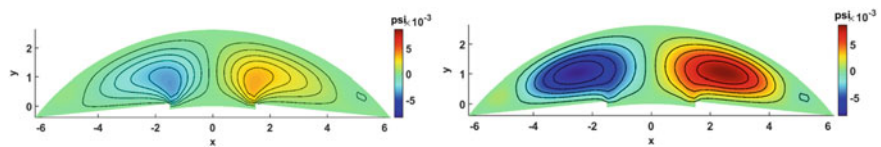


Fig. 5 *Supine position. Flow streamlines for $\theta_{amb} = 45\text{C}$ (Left) and $\theta_{amb} = 15\text{C}$ (Right)*

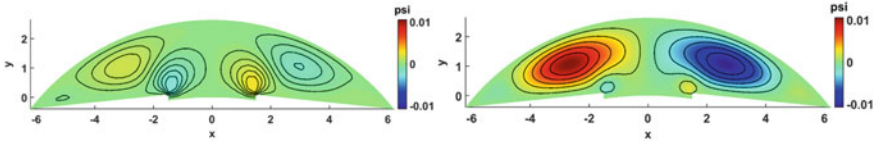


Fig. 6 Prone position. Flow streamlines for $\theta_{amb} = 45C$ (Left) and $\theta_{amb} = 15C$ (Right)

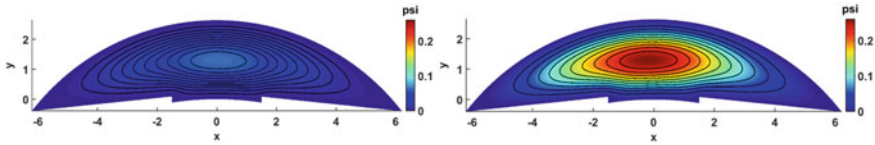


Fig. 7 Standing position. Flow streamlines for $\theta_{amb} = 45C$ (Left) and $\theta_{amb} = 15C$ (Right)

5 Conclusions

This work presents a numerical scheme that leverage operator splitting to maintain at the discrete level the physical properties that the solution enjoys at the continuous level. Our current observations concur with the velocity field as predicted by other simulation studies [13]. In particular, the maximum, minimum, and mean velocities are close to the predicted values. Previous analytical models of fluid flow in human eyes in different postures (supine and vertical) have concluded that the dominant mechanism that influences the fluid field in the anterior chamber is the convection-induced buoyancy flow. Similar observations have been made in a rabbit eye model [12]. Free convection is also claimed to be the major mechanism of deposition of cells/particles on the endothelial surface (e.g., formation of Krukenberg's spindle) [1]. Thus, the proposed numerical method has the potential to serve as a robust framework on which additional physical mechanisms can be incorporated.

References

1. Boushehrian, H.H., Abouali, O., Jafarpur, K., Ghaffarieh, A., Ahmadi, G.: Relationship between saccadic eye movements and formation of the Krukenberg's spindle—a CFD study. *Math. Med. Biol. J. IMA* **34**(3), 293–312 (2016)
2. Carichino, L., Guidoboni, G., Szopos, M.: Energy-based operator splitting approach for the time discretization of coupled systems of partial and ordinary differential equations for fluid flows: The stokes case. *J. Comput. Phys.* **364**, 235–256 (2018)
3. Cooke, E., Xia, Y.: Using fluid mechanics to optimise drug delivery in the eye. Technical report (2015)
4. Dvoriashyna, M., Pralits, J.O., Tweedy, J.H., Repetto, R.: Mathematical models of aqueous humor production, flow and drainage. In: Guidoboni, G., Harris, A., Sacco, R. (eds.) *Ocular Fluid Dynamics. Anatomy, Physiology, Imaging Techniques, and Mathematical Modeling*, Chap. 9, pp. 227–263. Springer-Birkhäuser (2020)

5. Geuzaine, C., Remacle, J.-F.: Gmsh: a 3-d finite element mesh generator with built-in pre-and post-processing facilities. *Int. J. Numer. Methods Eng.* **79**(11), 1309–1331 (2009)
6. Glowinski, R.: Numerical methods for fluids (part 3). *Handb. Numer. Anal.* **9**(3) (2003)
7. Guidoboni, G., Glowinski, R., Cavallini, N., Canic, S.: Stable loosely-coupled-type algorithm for fluid-structure interaction in blood flow. *J. Comput. Phys.* **228**(18), 6916–6937 (2009)
8. Hecht, F.: New development in freefem++. *J. Numer. Math.* **20**(3–4), 251–266 (2012)
9. Heys, J.J., Barocas, V.H.: A Boussinesq model of natural convection in the human eye and the formation of Krukenberg’s spindle. *Annals Biomed. Eng.* **30**(3), 392–401 (2002)
10. Kim, J., Kudisch, M., Mudumba, S., Asada, H., Aya-Shibuya, E., Bhisitkul, R.B., Desai, T.A.: Biocompatibility and pharmacokinetic analysis of an intracameral polycaprolactone drug delivery implant for glaucoma. *Investig. Ophthalmol. Vis. Sci.* **57**(10), 4341–4346 (2016)
11. Kinoshita, S., Koizumi, N., Ueno, M., Okumura, N., Imai, K., Tanaka, H., Yamamoto, Y., Nakamura, T., Inatomi, T., Bush, J., et al.: Injection of cultured cells with a rock inhibitor for bullous keratopathy. *N. Engl. J. Med.* **378**(11), 995–1003 (2018)
12. Kumar, S., Acharya, S., Beuerman, R., Palkama, A.: Numerical solution of ocular fluid dynamics in a rabbit eye: parametric effects. *Annals Biomed. Eng.* **34**(3), 530 (2006)
13. Ooi, E.-H., Ng, E.Y.-K.: Simulation of aqueous humor hydrodynamics in human eye heat transfer. *Comput. Biol. Med.* **38**(2), 252–262 (2008)
14. Sacco, R., Guidoboni, G., Mauri, A.G.: *A Comprehensive Physically-based Approach to Modeling in Bioengineering and Life Sciences*. Academic Press, Elsevier (2019)

Modelling Thermal Aspects of Decomposition



L. Calla and C. Sean Bohun

Abstract Temperature modelling at a crime scene is crucial for forensic investigators to estimate the minimum postmortem interval (PMI_{min}) of a cadaver. Upon death, insect species deposit eggs, and the resulting rate of development of the larvae are primarily temperature driven. By knowing the historical ambient temperature, development stage, and specific species of the larvae found on a cadaver, an accurate estimate can be made for this interval. The actual temperature of the remains prior to discovery cannot be determined without at least historical environmental temperature data. In this research we examine the possibility of inferring the thermal environment of the growing larvae and extracting a characteristic heat flux profile intrinsic to the growing insect population. The external environmental temperature can be combined with this profile to provide a much more accurate predictor of the temperature experienced by the larvae than is currently used and consequently a more reliable estimate of the PMI_{min} .

Keywords Forensic Entomology · Mathematical modelling · PMI

1 Introduction

Within the first two or three days after death, medical techniques can reliably estimate a PMI [3, 8]. Beyond this time period, forensic entomology provides a minimum postmortem interval (PMI_{min}) estimate by aging the immature insects feeding on the body. PMI_{min} describes the time between the first colonizing insects laying eggs on the body and the time of discovery of the remains. Insects are coldblooded; therefore their rate of development is affected by the temperature to which they are exposed. Each species requires a different total thermal threshold to complete

L. Calla (✉) · C. S. Bohun
University of Ontario Institute of Technology, Oshawa, Ontario, Canada
e-mail: leanna.calla@ontariotechu.net

C. S. Bohun
e-mail: sean.bohun@uoit.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_46

their development. Using this predictable metabolism, a thermal summation model is used to calculate when the first eggs were laid onto the decomposing remains. Daily temperature averages give yield to accumulated degree days (ADD), or more precise hourly temperatures would give yield to accumulated degree hours (ADH). Starting very early¹ in the decomposition of the cadaver, accurately predicting the process of decomposition depends on an accurate knowledge of the conditions experienced by insects as they develop. In part, this is due to the stages of the body's decay having a strong relationship with the succession of arthropods involved.

The succession of the insects at the scene and their rate of their development relies heavily on the ambient temperature [2, 11]. To predict the temperatures experienced by the arthropods, practitioners compute a linear regression to best match the measured temperature at the location of the cadaver with a nearby weather station. While this has the capability of giving an accurate prediction, especially over a long period of time, it neglects a number of known phenomena. It has been experimentally determined that the presence of larval masses elevate the cadaver temperature with large discrepancies of 10–20 °C above the ambient environmental temperature at the scene [2, 6, 9, 10]. Another noticeable discrepancy is delay in heating, which can be attributed to shelter from wind, sunlight or other natural elements at the body's location [6, 7].

2 Mathematical Model

To improve PMI_{min} predictions, forensic researchers require accurate estimates for the temperature profile which the insects experienced throughout their development. We created a mathematical model for heat transfer of the cadaver over time, which involves scene specific parameters. The heat transfer model will involve a parameter to represent the coupling with the environment, and a parameter for the intrinsic heat flux experienced by a cadaver.

We begin by considering case studies performed by the Forensic Ecology Research Facility at the University of Ontario Institute of Technology.² One study was performed in July 2016 and the other in September 2016, both of which were an outdoor pig decomposition study.³ In these experiments environmental and internal temperatures are recorded.

Figure 1 shows the temperature profiles for the first and last five days of the September study. The July study is omitted for brevity. In these images, the ambient environment temperature (green) is plotted with the temperature in the mouth cavity of the cadaver (red). We note the following characteristics:

¹ Insect species Calliphoridae and Muscidae colonize within the first few hours after death [5].

² The first study commenced on 08/09/2016 at 10:00 EST and terminated on 07/10/2016 at 23:59 EST and is referred to as the *September* study. The *July* study is slightly shorter, commencing on 05/07/2016 at 11:00 EST and 29/07/2016 at 23:59 EST.

³ According to [6], a domestic pig is the best representative in the field to emulate a human cadaver.

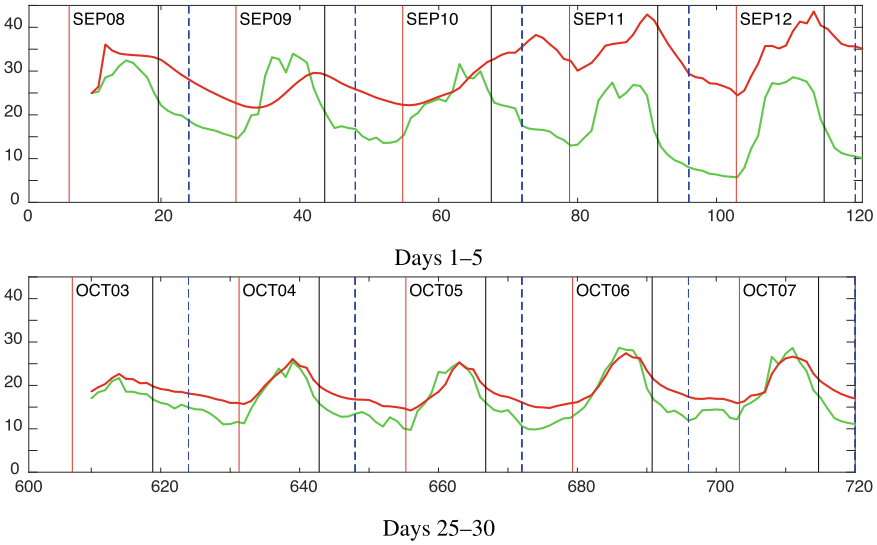


Fig. 1 Recorded temperatures of the environment (green) and the cadaver (red) for the September study. Vertical lines indicate (i) sunrise (red); (ii) sunset (black); and (iii) midnight (dashed blue). Temperatures are in °C and time is measured in hours since 08/09/2016, 00:00 EST

- The average internal logger is often warmer than the average environment logger.
- The difference in average temperatures between the two curves decays over time.
- In the first five days, the internal logger experiences less variation between the maximums and minimums than the environment logger.
- There is a significant initial phase shift between the internal and environment logger.
- Day and night behaviour is distinct for both internal and external temperature profiles.

These collective behaviours suggest that an averaged thermal model for the cadaver, that is driven by an external temperature, may be sufficient to reproduce these properties.

Within the cadaver domain denoted by Ω , we suppose that the temperature experienced by the insects, $T(\mathbf{x}, t)$, satisfies a heat equation with heat sources/sinks. That is,

$$\rho c_p \frac{\partial T}{\partial t} = k \nabla^2 T + q(\mathbf{x}, t), \quad \mathbf{x} \in \Omega, \quad t > 0, \quad (1a)$$

where the mass density ρ , specific heat c_p , and thermal conductivity k are all assumed to be constant for convenience. Through the boundary of Ω , denoted by $\partial\Omega$, the heat flux is taken to be proportional to the temperature difference across this surface. In particular

$$-k \left. \frac{\partial T}{\partial \mathbf{n}} \right|_{\partial \Omega} = h(T - T_0), \quad \mathbf{x} \in \partial \Omega, \quad t > 0, \quad (1b)$$

where $\hat{\mathbf{n}}$ denotes the outward unit normal vector at the surface, $T_0(t)$ is the external temperature at time t , and h is a constant effective heat transfer coefficient. Rather than solving for the temperature at every point within the cadaver, we define an average temperature within Ω as

$$\bar{T}(t) = \frac{1}{|\Omega|} \int_{\Omega} T(\mathbf{x}, t) \, d\mathbf{x} \quad (2)$$

where $|\Omega|$ denotes the volume of the domain. As well as these assumptions, we choose a scaling that takes into consideration a natural temperature scale of $[T_{\min}, T_{\max}]$, $\Delta T = T_{\max} - T_{\min}$ and time scale of τ (based on 12-h) so that

$$\bar{T} = T_{\min} + \Delta T \theta, \quad T_0 = T_{\min} + \Delta T \theta_0, \quad t = \tau \tilde{t}. \quad (3)$$

After dropping the tilde notation, the equation for the averaged temperature in the rescaled quantities becomes

$$\frac{d\theta}{dt} = c(t)(\theta_0(t) - \theta) + s(t), \quad (4a)$$

with

$$c(t) = \frac{h|\partial \Omega| \tau}{\rho c_p |\Omega|}, \quad s(t) = \frac{Q(t) \tau}{\rho c_p \Delta T}, \quad Q(t) = \frac{1}{|\Omega|} \int_{\Omega} q(\mathbf{x}, t) \, d\mathbf{x}, \quad (4b)$$

reducing to two, possibly time dependent, characteristic quantities $c(t)$ and $s(t)$.⁴

The parameter $c(t) \geq 0$ is a ratio of the thermal energy that passes through the boundary in the characteristic time to the thermal energy contained within the volume. In contrast, $s(t)$ models the energy density of a heat source ($Q(t) > 0$) or sink ($Q(t) < 0$) with respect to the characteristic thermal energy density within the body given by $\rho c_p \Delta T$. In [4] the cadaver is described throughout decomposition as deflating due to loss of mass and liquefying of internal organs. This supports the notion that c will slowly increase as the surface area to volume ratio of the cadaver increases provided that the heat transfer coefficient does not change. The variation of s is expected to follow the circadian rhythm of insect activity.

We consider (4a) as an inverse model to determine the structure of $c(t)$ and $s(t)$ so that $\theta(t)$ best matches the experimentally measured temperature of a thermal decomposition study. These parameters can depend on a variety of environmental factors and growth stages of the entomological species that are present. The power of

⁴ Choosing nominal values of $h \sim 1 \, \text{J s}^{-1} \text{m}^{-2} \text{K}^{-1}$, $\rho c_p \sim 10^6 \, \text{J m}^{-3} \text{K}^{-1}$, $\tau = 12 \times 3600 \, \text{s}$ and an effective cadaver thickness of $|\Omega|/|\partial \Omega| \sim 10^{-1} \, \text{m}$, gives an estimate of $c \sim 0.43$. Similarly with $\Delta T \sim 10 \, \text{K}$ and $Q \sim 10^4 \, \text{J s}^{-1} \text{m}^{-3}$ yields $s \sim 43$.

the model, and its various approximations, lies in the notion that once a characteristic behaviour of $s(t)$ and $c(t)$ is known, expression (4a) acts as a predictive tool for the thermal evolution of the body temperature, with only the environmental temperature history as its input.

Some insight into the structure of the solution can be obtained by writing $\theta = u + v$ where u and v satisfy

$$\frac{du}{dt} = c(t)(\theta_0(t) - u), \quad u(\tau_0) = \theta(\tau_0); \quad \frac{dv}{dt} = -c(t)v + s(t), \quad v(\tau_0) = 0. \quad (5a)$$

We can then solve for both $u(t)$ and $v(t)$ to obtain

$$\mu(t)u(t) = \theta(\tau_0) + \int_{\tau_0}^t \mu(\eta)c(\eta)\theta_0(\eta) d\eta, \quad \mu(t)v(t) = \int_{\tau_0}^t \mu(\eta)s(\eta) d\eta \quad (5b)$$

where

$$\mu(t) = \exp\left(\int_{\tau_0}^t c(\eta) d\eta\right). \quad (5c)$$

Recombining the two expressions gives the solution as

$$\theta(t) = u(t; c) + \frac{1}{\mu(t)} \int_{\tau_0}^t \mu(\eta)s(\eta)d\eta, \quad (6)$$

illustrating that the external temperature, modulated by the effective heat transfer coefficient, acts as a separate external heat flux.

3 Model Implementation

We consider an implementation where the time interval is broken into a disjoint union of $N + 1$ intervals, $\{I_j = [\tau_j, \tau_{j+1})\}_{j=0}^N$ divided at the sunrise/sunset times. On each interval I_j we suppose a constant external flux s_j and construct the vector $\mathbf{s} = (s_0, \dots, s_N)^T$ for the entire interval $[\tau_0, \tau_{N+1}) = \cup_{j=0}^N I_j$. For this entire interval we take a constant value for $c(t)$ so that $c(t) = c_0 > 0$, $\tau_0 \leq t < \tau_{N+1}$. In this case the solution takes the form

$$\begin{aligned} \theta(t; c_0, \mathbf{s}) &= \theta(t_0)e^{-c_0(t-\tau_0)} + c_0 \int_{\tau_0}^t e^{-c_0(t-\eta)}\theta_0(\eta) d\eta \\ &+ \frac{s_l}{c_0} (1 - e^{-c_0(t-\tau_l)}) + \frac{1}{c_0} \sum_{j=0}^{l-1} s_j (e^{-c_0(t-\tau_{j+1})} - e^{-c_0(t-\tau_j)}), \quad (7) \end{aligned}$$

for $\tau_l \leq t < \tau_{l+1}$, and $l = 0, 1, \dots, N$. Using this approximate model as our working platform, we now attempt to determine an external coupling, c_0 , and characteristic flux, \mathbf{s} , that describe a given thermal decomposition study.

To test this model, we consider a thermal decomposition experiment over a number of weeks and measure both the internal and external temperature periodically resulting in $M \gg N + 1$ samples, $\{\theta_j^{\text{meas}}\}_{j=1}^M$ and $\{\theta_{0,j}^{\text{meas}}\}_{j=1}^M$ taken at $\{t_j\}_{j=1}^M$ respectively. Writing (7) as a linear system in \mathbf{s} , ideally c_0 and \mathbf{s} would be chosen so that

$$f(c_0, \mathbf{s}) = \|\mathbf{u}(c_0) - \hat{\theta} + B(c_0)\mathbf{s}\|_{\ell^2(\mathbb{R}^M)}^2 \tag{8a}$$

is minimized where $\hat{\theta} = (\theta_1^{\text{meas}}, \dots, \theta_M^{\text{meas}})^\top$, $\mathbf{u}(c_0) = (u(t_1; c_0), \dots, u(t_M; c_0))^\top$, and for $t_i \in I_l$

$$[B]_{ij} = \frac{1}{c_0} \begin{cases} e^{-c_0(t_i - \tau_j)} - e^{-c_0(t_i - \tau_{j-1})}, & 1 \leq j \leq l, \\ 1 - e^{-c_0(t_i - \tau_l)}, & j = l + 1, \\ 0, & l + 2 \leq j \leq N + 1. \end{cases} \tag{8b}$$

A standard calculation to find the minimum with respect to \mathbf{s} gives

$$\mathbf{s}^* = \arg \min_{\mathbf{s} \in \mathbb{R}^{N+1}} f(c_0, \mathbf{s}) = (B^\top B)^{-1} B^\top (\hat{\theta} - \mathbf{u}) \tag{9}$$

leaving a nonlinear, one-dimensional minimization problem to find

$$c_0^* = \arg \min_{c_0 > 0} f(c_0, \mathbf{s}^*) = \arg \min_{c_0 > 0} \left\| (\mathbf{I} - B(B^\top B)^{-1} B^\top) (\mathbf{u} - \hat{\theta}) \right\|_{\ell^2(\mathbb{R}^M)}^2. \tag{10}$$

Using an alternative norm may lead to different results and this will be explored at a later date.

4 Results

The approximate model (7) is applied to the two independent thermal decomposition studies. We use these studies as a ground truth to find the coupling coefficient and heat flux parameters.

4.1 Analysis of the Coupling Coefficient

We begin by finding the best coupling coefficient that minimizes the percent relative error defined as

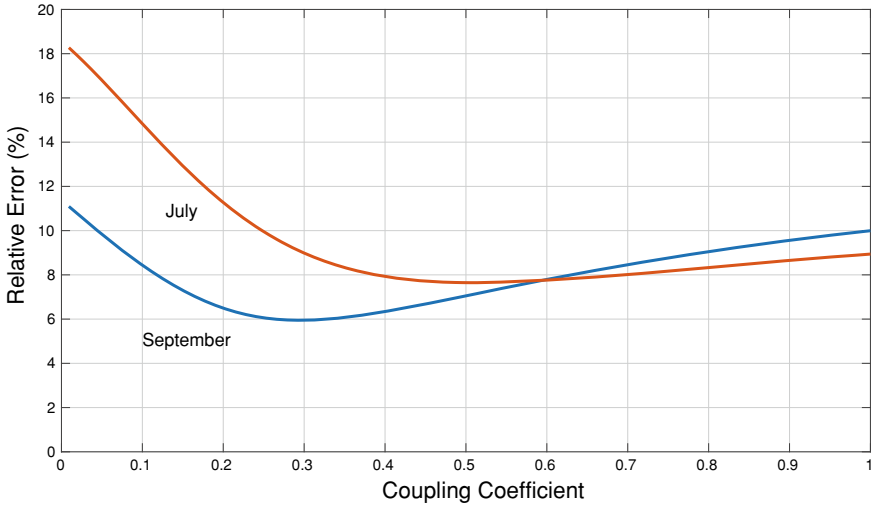


Fig. 2 The relative error defined by expression (11) as a function of c_0 for the duration of each decomposition study for the July ($c_0^* \simeq 0.51$) and the September ($c_0^* \simeq 0.24$) experiments

$$\text{Percent relative error} = \frac{1}{\|\hat{\theta}\|_{\ell^2}} \left\| (\mathbf{I} - B(B^T B)^{-1} B^T) (\mathbf{u} - \hat{\theta}) \right\|_{\ell^2} \times 100 \quad (11)$$

which is illustrated in Fig. 2, assuming a single value of c_0 for the entire interval. There are well defined minimal relative errors obtained for both studies. The value of $c_0^* \simeq 0.51$ for July and $c_0^* \simeq 0.24$ for September.

We have assumed that c is constant, and to test this assumption, we find the optimal value c_0^* as a function of the duration of the experiment. Figure 3 illustrates that c_0^* increases with the duration, but the $\pm 1\%$ relative error bounds are quite large. This spread (in either experiment) indicates the predictability of the model. Consistent with Fig. 2, the fit is not very sensitive to the value of c_0^* .

4.2 Temperature Prediction and Heat Flux Analysis

With an optimal choice of coupling coefficient, c_0^* , and the corresponding heat flux vector, \mathbf{s}^* , we may return to expression (7) to estimate the cadaver temperature. Figure 4 shows this prediction for the September study. The fit is better in the last five days. A possible reason is indicated in Fig. 3 which indicates changes in c may be aligned with the decomposition regime. This is especially noticeable in the July study.

Figure 5 shows our findings for the heat flux of each study. The vertical lines indicate progression through the decomposition regimes with the September study failing to reach dry remains. The forensics literature of this process is anecdotal, with

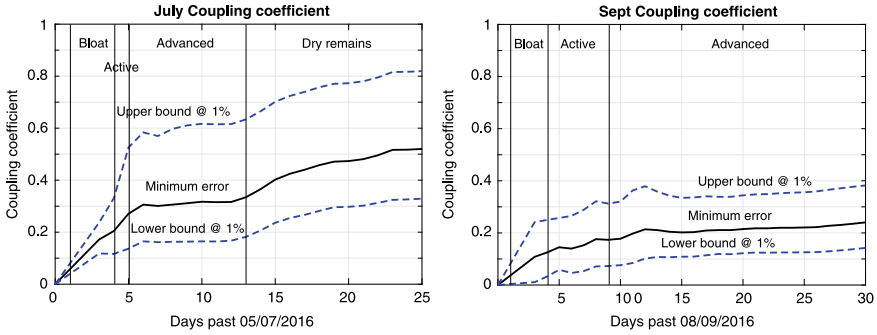


Fig. 3 The optimal coupling coefficient, c_0^* , as a function of the number of days used in each study with the decomposition regimes indicated. The upper and lower dashed curves indicate the location of $\pm 1\%$ relative error away from the optimal value

Anderson noting increased heating in the first few days with pig carcasses, followed by a rapid drop after a few weeks [4]. It has also been noticed that colonization of a cadaver follows a specific pattern; with the early stages of bloat and decay experiencing the most variety, and largest quantities of insect species [3]. In Fig. 5,

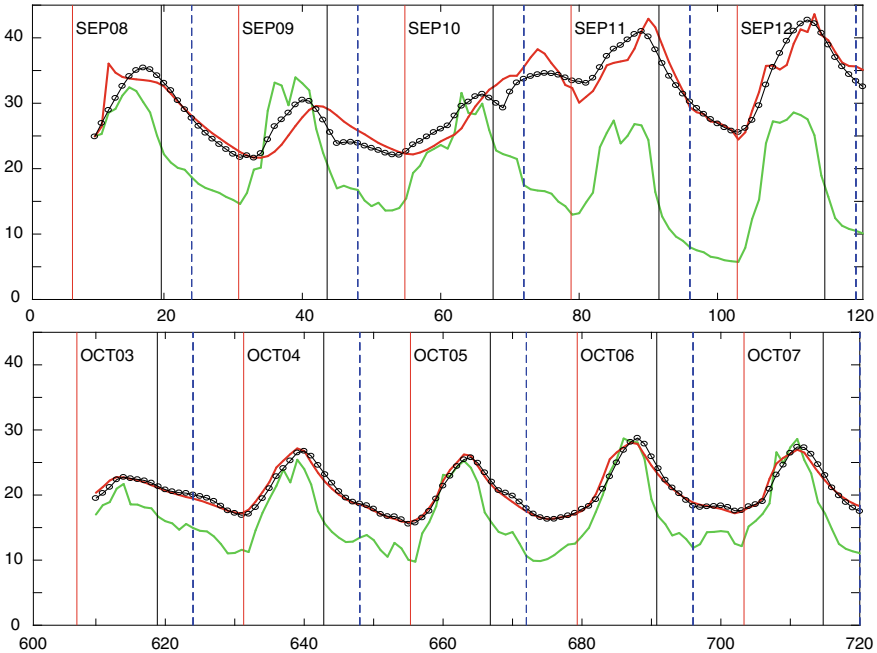


Fig. 4 The model prediction of the temperature (dotted black) for the first five and last five days of the September study. For reference, the data in Fig. 1 is also included

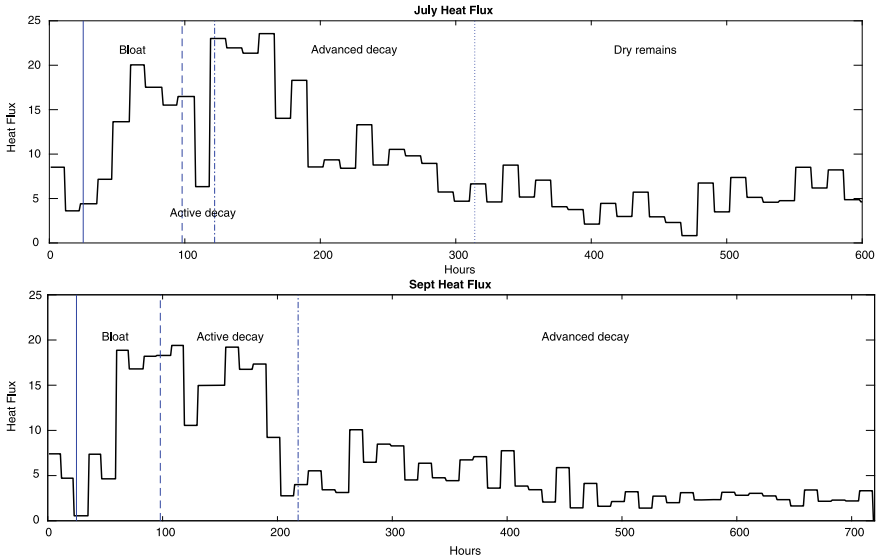


Fig. 5 The predicted piece-wise constant heat flux, s^* , for the duration of the July and September studies. The characteristics of the two curves are similar with two initial broad peaks followed by a decay to the background level

a characteristic pattern in the heat flux is indicated, independent of the season, July or September. There are two broad initial bursts of activity followed by a residual background flux which is consistent with the observations described in [1].

The heating of the cadaver is driven by both environmental effects as well as the presence of insects. An accurate prediction of the cadaver temperature is a key component of accurately determining the time of death in that it modulates the growth rate of colonizing insects. This research has used thermal decomposition studies to extract both an intrinsic heat flux for the colonizing insect species, s^* , and an effective coupling to the environment, c_0^* . With these parameters, the model (7) can rapidly predict the cadaver temperature much more accurately than current convention of using linear regression with the environmental temperature. This accuracy is reflected in a more reliable estimate of time of death.

5 Conclusion

In this research we attempted to extract an intrinsic heat response from colonizing insect species and a single coupling coefficient in an application of an inverse model corresponding to (4a). It is expected that s^* is a function of the specific colonizing species and that the coupling c_0^* is based on a variety of environmental factors. Armed with appropriate $s(t)$ and $c(t)$, (4a) provides an efficient and rapid predictor

of a cadaver temperature as it is driven by an external temperature source. In effect, this is a novel, non-intrusive way to accurately predict the thermal environment of the colonizing insects and will allow for increased accuracy in postmortem interval predictions.

6 Future Work

A single effective value of c_0^* was chosen for simplicity but a reanalysis of the data with $c(t)$ changing at the same frequency as $s(t)$ is being undertaken. The authors agree with an anonymous referee that this is an essential first step in further developing this method. The variation of c_0 with season and geographic location is also currently under study.

Acknowledgements Thank you to H. LeBlanc and A. Skopyk for their case study data and expertise in forensic entomology.

References

1. Alvers, M., Soares, J.: Diurnal variation of soil heat flux at an Antarctic local area during warmer months. *Appl. Environ. Soil Sci.* (2016)
2. Amendt, J., Campobasso, C., Gaudry, E., Reiter, C., LeBlanc, H., Hall, M.: Best practice in forensic entomology—standards and guidelines. *Int. J. Legal Med.* **121**, 90–104 (2007)
3. Amendt, J., Krettek, R., Zehner, R.: Forensic entomology. *Naturwissenschaften* (2004)
4. Anderson, G., VanLaerhover, S.: Initial studies on insect succession on carrion in Southwestern British Columbia. *J. Forensic Sci.* **41**, 617–625 (1996)
5. Campobasso, C., Di Vella, G., Introna, F.: Factors affecting decomposition and diptera colonization. *Forensic Sci. Int.* **120**, 18–27 (2001)
6. Catts, E., Goff, M.: Forensic entomology in criminal investigations. *Annual Rev. Entomol.* **37**, 253–272 (1992)
7. Charabidze, D., Hedouin, V.: Temperature: the weak point of forensic entomology. *Int. J. Legal Med.* **133**, 633–639 (2019)
8. Clark, M.A., Worrell, M.B., Pless, J.E.: Postmortem changes in soft tissues. In: *Forensic Taphonomy: The Postmortem Fate of Human Remains*, pp. 151–164 (1997)
9. Hofer, I., Hart, A., Martin-Vega, D., Hall, M.: Optimising crime scene temperature collection for forensic entomology casework. *Forensic Sci. Int.* **270**, 129–138 (2017)
10. Slone, D., Gruner, S.: Thermoregulation in larval aggregations of carrion-feeding blow flies (diptera: Calliphoridae). *J. Med. Entomol.* **44**(3), 516–523 (2007)
11. Tomberlin, J., Mohr, R., Benbow, A., VanLaerhoven, S.: A roadmap for bridging basic and applied research in forensic entomology. *Annual Rev. Entomol.* **56**, 401–421 (2011)

Mathematical Modeling of the Steady-State Behavior of Nitric Oxide in Brain



Corina S. Drapaca and Andrew Tamis

Abstract Nitric oxide (NO) is a small diffusible molecule that plays an important role in brain's signalling processes and regulation of cerebral blood flow and pressure. While most of the NO production is achieved through various chemical reactions taking place in the neurons, endothelial cells, and red blood cells, only the endothelial NO is activated by the shear stress at the blood-endothelium interface. NO is removed from the brain by blood's hemoglobin and through diffusion and other chemical processes. Given its relevance to brain functions, numerous studies on NO exist in the literature. The majority of the mathematical models of NO biotransport are diffusion-reaction equations predicting the spatio-temporal distribution of NO concentration either inside or outside the blood vessels, and do not account for the endothelial NO production through mechanotransduction. In this paper we propose a mathematical model of the steady-state behavior of NO in the brain that links the NO synthesis and inactivation from inside and outside a cerebral arteriole and the blood flow. The blood flow is assumed to be a Poiseuille flow, and we use two models of blood: viscous Newtonian and non-local non-Newtonian fluids. The model is used to study through numerical simulations the effects of the cerebral blood pressure on the NO concentration.

Keywords Cerebral Nitric Oxide · Poiseuille Flow · Mechanotransduction · Fractional Calculus · Nonlocality

1 Introduction

Nitric oxide (NO) is a free radical gas involved in many critical bio-chemical processes taking place in living organisms. In particular, NO acts as a neuro-glial-vascular messenger and regulator of cerebral blood flow [1, 2, 6, 7, 10, 11, 14,

C. S. Drapaca (✉) · A. Tamis
Pennsylvania State University, State College, University Park, PA 16802, USA
e-mail: csd12@psu.edu

A. Tamis
e-mail: apt5288@psu.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_47

511

15, 18, 20–22]. NO is produced by synthesis reactions taking place in specific neurons [10], vascular endothelium and choroid plexus [11], and red blood cells [15]. While neuronal NO is involved in learning, memory formation, and regulation of the cerebral blood flow [11], the endothelial NO maintains cerebral microcirculation by guiding vasomotor responses and vasoprotection processes, and by reducing the cerebral blood pressure [1, 10, 12, 13]. The endothelial NO can also be produced through mechanotransduction initiated by the shear stress at the endothelium-blood interface [25]. The NO formed in the deoxygenated red blood cells is involved in the red blood cells deformability [3]. NO is removed from the brain by blood's hemoglobin and through diffusion and other chemical processes [15].

Given the essential role played by NO in various physiological and pathophysiological processes, especially those involved in brain functions, numerous mathematical models of NO spatio-temporal dynamics exist in the literature [5, 14, 16, 26, 27] (a comprehensive review of mathematical models of NO biotransport is given in [5]). The majority of these models are reaction-diffusion equations describing NO syntheses (production) and inactivation (loss) either inside or outside the blood vessels. The shear-induced production of the endothelial NO and the coupling of the NO contributions from the inside and outside of the blood vessels are usually modelled as known boundary conditions. Since these models do not incorporate any mechanical properties of cells and their interactions, they cannot accurately predict the NO effects on, for instance, cerebral blood flow and vasculature, and ultimately on brain functions. A mathematical model that couples the NO spatio-temporal dynamics and the mechanical behaviors of blood flow and vasculature could prove essential in the development of successful NO-based therapies for clinical conditions associated with disturbances in NO production and/or signaling [4].

In this paper we propose a mathematical model for the steady-state behavior of NO in brain that incorporates the production of endothelial NO through shear-induced mechanotransduction. The steady-state reaction-diffusion equation for the NO concentration includes production and decay terms from the inside and outside of a cerebral arteriole which are taken from [14, 16, 26, 27]. A new NO production term is added to the equation that models the shear-induced mechanotransduction of NO in endothelium. This production term is assumed to be proportional to the concentration of NO which is in agreement with the experimental observations reported in [25]. We further conjecture that the production rate of this term is proportional to the viscous dissipation at the blood flow-endothelium interface. Dissipation is calculated from assumptions on the mechanical properties of the blood flow and arteriolar wall. For now, the wall is assumed to be rigid and permeable only to NO. The blood flow is modelled as a Poiseuille flow, and we use two models of blood: viscous Newtonian and non-local non-Newtonian fluids. The non-Newtonian nature of blood becomes apparent in the smaller vessels such as the arterioles. Here we use the non-local non-Newtonian model of blood proposed in [8] which accounts for long-range chemo-mechanical interactions of the red blood cells *in vivo*. The model is used to study through numerical simulations the effects of the cerebral blood pressure on the NO concentration.

The paper is structured as follows. Section 2 presents the mathematical model, while the numerical results are shown in Sect. 3. The paper ends with conclusions and suggestions for future work.

2 Mathematical Model

The geometric domain shown in Fig. 1 is made of concentric horizontal axis-symmetric circular cylinders. The cylinders have rigid walls which are permeable only to NO. The blood flows through the lumen region of radius R . The endothelium layer of the arteriole has thickness h and is considered separately from the other arteriolar layers because it is a NO production site. The next region of thickness d is made of the other vascular layers and extracellular space. Lastly, the region of thickness g represents a group of neurons that produce NO.

As in [26, 27], the NO transport by convection is neglected. The blood is an incompressible fluid in a three-dimensional fully-developed steady laminar flow. The flow is axis-symmetric and driven by an externally applied pressure gradient. No body forces are acting on the blood. In cylindrical coordinates (r, θ, z) , only the axial component of the blood's velocity is non-zero. Thus, at steady-state, the concentration of NO, c_{NO} , and the axial component of the blood's velocity, w , depend only on the independent variable r .

The NO is produced in the region $[R, R + h] \cup [R + h + d, R + h + d + g]$ and decays in the region $[0, R] \cup [R + h, R + h + d + g]$. In addition, NO diffuses radially on $[0, R + h + d + g]$. Thus, the balance law of mass at steady-state in cylindrical coordinates is:

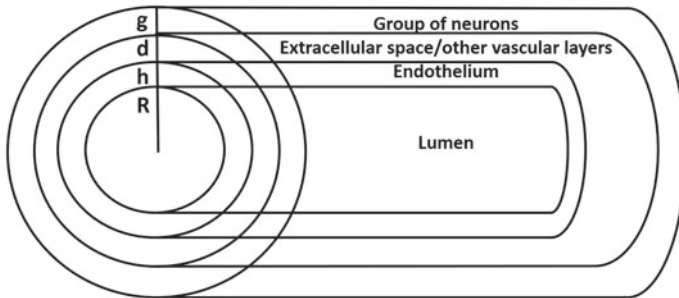


Fig. 1 The geometric domain is made of concentric horizontal axis-symmetric circular cylinders. The regions correspond to the lumen of radius R , endothelium of thickness h , a region of thickness d made of other arteriolar layers and extracellular space, and a region of thickness g filled with neurons

$$D_{NO} \left(\frac{d^2 c_{NO}}{dr^2} + \frac{1}{r} \frac{dc_{NO}}{dr} \right) + v_1 H(r - (R + h + d)) - \frac{V_{max} c_{NO}}{K_{max} + c_{NO}} H(r - (R + h)) + \frac{\sigma_{rz}}{\tau_w} \frac{dw}{dr} (H(r - R) - H(r - (R + h))) c_{NO} - \lambda (1 - H(r - R)) c_{NO} = 0 \quad (1)$$

In Eq. (1), D_{NO} is the diffusion coefficient of NO, v_1 is the constant rate of NO synthesis in the neurons, V_{max} is the maximum rate at saturating concentration of NO in the region $[R + h, R + h + d + g]$, and K_{max} is the NO concentration at which the reaction rate is $V_{max}/2$ in the region $[R + h, R + h + d + g]$ [14]. The viscous dissipation is the product between blood’s shear stress σ_{rz} and $\frac{dw}{dr}$. The shear stress at the lumen-endothelium interface is $\tau_w = \sigma_{rz}(R)$. Lastly, the decay of NO due to the hemoglobin in blood is assumed to happen at a constant rate λ [16, 26, 27]. The Heaviside step function is denoted by H . The newly introduced fourth term in Eq. (1) is *the shear-induced production of NO in the endothelium*. Equation (1) was solved numerically and thus the presence of the discontinuous Heaviside step function in the equation did not pose any issues. In subsequent work a mathematical analysis of this equation with smooth coefficients will be performed to better understand the mathematical behavior of the solution and its physical interpretation.

The expressions for σ_{rz} , $\frac{dw}{dr}$, and τ_w are given further. The following assumptions specific to a Poiseuille flow in the region $[0, R]$ are made. Let $C = \frac{dp}{dz} < 0$ be the constant, externally applied pressure gradient, where $p(z)$ is the hydrostatic pressure of blood. The lumen-endothelium interface is assumed to be a no slip boundary. The boundary condition at $r = 0$ expresses the axial symmetry of the flow.

Newtonian Model of Blood For an incompressible viscous Newtonian fluid, the shear stress σ_{rz} and shear rate $\frac{dw}{dr}$ are related through the following constitutive relation:

$$\sigma_{rz} = \mu \frac{dw}{dr} \quad (2)$$

where μ is the dynamic viscosity. The solution of the Navier-Stokes equations at equilibrium with boundary conditions:

$$w(R) = \frac{dw}{dr}(0) = 0 \quad (3)$$

is:

$$w(r) = \frac{C}{4\mu} (r^2 - R^2) \quad (4)$$

Thus:

$$\frac{dw}{dr} = \frac{Cr}{2\mu}, \sigma_{rz} = \frac{Cr}{2}, \tau_w = \frac{CR}{2} \tag{5}$$

Since $C < 0$, a change of sign is required in front of the fourth term of equation (1) such that the term models NO production. Thus, by replacing formulas (5) in Eq. (1), Eq. (1) becomes:

$$D_{NO} \left(\frac{d^2 c_{NO}}{dr^2} + \frac{1}{r} \frac{dc_{NO}}{dr} \right) + v_1 H(r - (R + h + d)) - \frac{V_{max} c_{NO}}{K_{max} + c_{NO}} H(r - (R + h)) - \frac{Cr^2}{2\mu R} (H(r - R) - H(r - (R + h))) c_{NO} - \lambda(1 - H(r - R)) c_{NO} = 0 \tag{6}$$

Non-local Non-Newtonian Model of Blood The shear stress σ_{rz} for an incompressible non-local non-Newtonian fluid is given by [8]:

$$\sigma_{rz} = \mu D_r^\alpha w(r) \tag{7}$$

where $D_r^\alpha w(r)$ is the left-sided Caputo fractional derivative of order α which, by definition, is:

$$D_r^\alpha w(r) = \frac{1}{\Gamma(m - \alpha)} \int_0^r \frac{1}{(r - s)^{\alpha+1-m}} \frac{d^m w(s)}{ds^m} ds, \quad m - 1 < \alpha < m$$

or

$$D_r^\alpha w(r) = \frac{d^m}{dr^m} w(r), \quad \alpha = m$$

for $m \in \{1, 2, 3, \dots\}$. From a physical point of view, $D_r^\alpha w(r)$ in formula (7) represents the shear rate of order α .

For $\alpha = 1$, formula (7) reduces to formula (2) and the physical parameter μ becomes the apparent viscosity. Parameter $\alpha \neq 1$ gives an intrinsic coupling between flow and the continuous rearrangement of the fluid’s microstructure during flow. The information about this coupling is lost when $\alpha = 1$. Thus, the constitutive equation (7) for $\alpha \neq 1$ models long-range (non-local) interactions among cells caused by and contributing to blood flow in vivo.

The solution of the Navier-Stokes equations at equilibrium with the boundary conditions:

$$w(R) = \frac{d^k}{dr^k} w(0+) = 0, \quad k = 1, 2, \dots, m - 1, \quad m - 1 < \alpha < m \tag{8}$$

is [8]:

$$w(r) = \frac{C}{2\mu\alpha(\alpha + 1)\Gamma(\alpha)} (r^{\alpha+1} - R^{\alpha+1}). \tag{9}$$

Thus:

$$\frac{dw}{dr} = \frac{Cr^\alpha}{2\mu\alpha\Gamma(\alpha)}, \sigma_{rz} = \frac{Cr}{2}, \tau_w = \frac{CR}{2} \tag{10}$$

Lastly, by replacing formulas (10) in Eq. (1) and using the same sign convention as before, the following equation is obtained for c_{NO} :

$$D_{NO} \left(\frac{d^2c_{NO}}{dr^2} + \frac{1}{r} \frac{dc_{NO}}{dr} \right) + v_1 H(r - (R + h + d)) - \frac{V_{max}c_{NO}}{K_{max} + c_{NO}} H(r - (R + h)) - \frac{Cr^{\alpha+1}}{2\mu R\alpha\Gamma(\alpha)} (H(r - R) - H(r - (R + h)))c_{NO} - \lambda(1 - H(r - R))c_{NO} = 0 \tag{11}$$

If $\alpha = 1$, Eq. (11) reduces to Eq. (6). Thus, it suffices to build a numerical solution only for Eq. (11).

3 Results

The values of the parameters used in the numerical simulations are given in Table 1.

Two values for C are used which are named healthy and high. For the Newtonian model of blood, healthy and high pulse pressures in humans are estimated from [24]. For the non-local non-Newtonian model of blood, the healthy pressure gradient is

Table 1 List of parameters with corresponding values and units

Model of blood	Parameters	Values and units (Reference)
	R	25×10^{-6} m [16]
	h	0.5×10^{-6} m [16]
	d	4×10^{-6} m
	g	5×10^{-6} m [14]
	D_{NO}	3.3×10^{-9} m ² /s [14, 16]
	v_1	1.6×10^{-3} mol/(m ³ × s) [14]
	V_{max}	2×10^{-3} mol/(m ³ × s) [14]
	K_{max}	10^{-6} mol/m ³ [14]
	λ	2.3×10^2 1/s [26]
Newtonian	μ	3 g/(m × s) [17]
	C (healthy)	-5.3×10^8 g/(m ² × s ²) [24]
	C (high)	-7.9×10^8 g/(m ² × s ²) [24]
Non-local Non-Newtonian	α	1.97 [9]
	μ	$0.021 \times 10^{6(\alpha-2)}$ g/(m ^{2-α} × s) [9]
	C (healthy)	-7.122×10^7 g/(m ² × s ²) [9]
	C (high)	-7.122×10^{10} g/(m ² × s ²)

estimated from experiments performed on a living mouse and reported in [19]. A high pressure gradient in living mice was not found in the literature and, therefore, this is chosen to be a value at which a difference in c_{NO} is observed from the healthy case. However, this high value of C might not be physiological. Lastly, parameters α and μ of the non-local non-Newtonian model were found in [9] by fitting the speed given in formula (9) to the blood speed measured in vivo in a venule of a mouse cremaster muscle before systemic hemodilution [19].

Equation (11) is solved numerically using a zero Neumann boundary condition at $r = 0$:

$$\frac{dc_{NO}}{dr}(0) = 0 \tag{12}$$

and the following Dirichlet boundary condition at $r = R + h + d + g$ estimated from [14]:

$$c_{NO}(R + h + d + g) = 10^{-8} \text{ [mol/m}^3\text{]} \tag{13}$$

Numerical solutions are obtained using the built-in function **bvp5c** in MATLAB. The function **bvp5c** solves boundary-value problems for systems of first order ordinary differential equations using the four-stage Lobatto IIIA formula represented as an implicit Runge-Kutta formula [23]. The system of first order differential equations corresponding to Eq. (11) is:

$$\begin{aligned} \frac{dc_{NO}}{dr} &= s_{NO}, \\ \frac{ds_{NO}}{dr} &= -\frac{1}{r}s_{NO} - \frac{v_1}{D_{NO}}H(r - (R + h + d)) + \frac{V_{max}c_{NO}}{D_{NO}(K_{max} + c_{NO})}H(r - (R + h)) \\ &+ \frac{Cr^{\alpha+1}}{2D_{NO}\mu R\alpha\Gamma(\alpha)}(H(r - R) - H(r - (R + h)))c_{NO} + \frac{\lambda}{D_{NO}}(1 - H(r - R))c_{NO} \end{aligned} \tag{14}$$

Thus, the function **bvp5c** solves system (14) with boundary conditions (12)–(13) for the unknowns c_{NO} and s_{NO} .

The results are shown in Fig. 2. Both models of blood show similar profiles for c_{NO} for their respective healthy and high values of C . The profiles of c_{NO} inside and outside the arteriole agree with those found in the literature. However, in the endothelium region (Fig. 2c, d) the concentration of NO for the high value of C is slightly higher than the one corresponding to the healthy value of C . These findings suggest that the blood pressure could affect the concentration of NO. The concentration of NO is more sensitive to changes in C when using the Newtonian model of blood than when the blood is modeled as a non-local non-Newtonian fluid. Since the blood flowing through small arterioles is non-Newtonian, it is possible that the concentration of NO is not affected by higher values of the cerebral blood pressure which are within physiological limits. Nevertheless, given the uncertainties in the values of the parameters in Table 1 and the inconsistencies among these parameters (some parameters were estimated from slices of rat brains [14], others from cremaster muscles of mice [19], and the rest from humans [17, 24]), a careful sensitivity analysis needs to

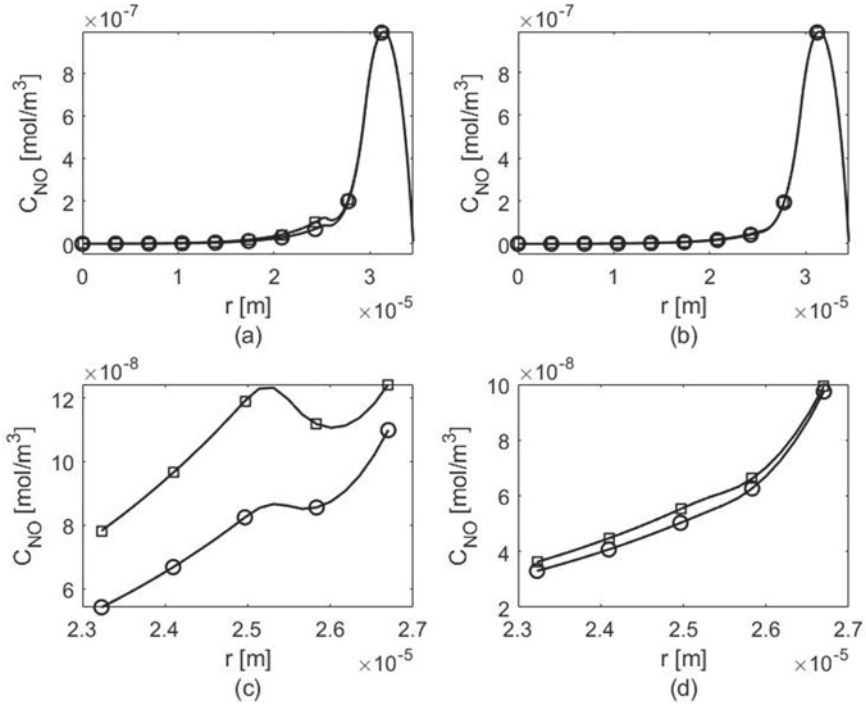


Fig. 2 Concentrations of NO for healthy (circle symbol) and high (square symbol) pressure gradients C : **a** Newtonian model of blood ($\alpha = 1$), **b** Non-local non-Newtonian model of blood ($\alpha = 1.97$). Zoom-ins of the plots **(a)** and **(b)** are shown in plots **(c)** and respectively **(d)**. The region around the endothelium is chosen for the zoom-ins. c_{NO} in the endothelium is higher for a higher value of C , as expected from Eqs. (6) and (11)

be performed in order to get a better understanding of the relationship between the NO concentration and the cerebral blood flow and pressure and confirm the validity of these preliminary results.

4 Conclusion

In this paper, a mathematical model was proposed to describe the steady-state behavior of NO in brain. The model is a one-dimensional, steady-state reaction-diffusion equation for the concentration of NO that includes production and decay terms from the inside and outside of a cerebral arteriole which were taken from the literature. A new production term is added to the equation that models the NO production in the endothelium through shear-induced mechanotransduction. This production term is proportional to the concentration of NO and the corresponding production rate is

proportional to the viscous dissipation at the blood flow-endothelium interface. The dissipation was calculated using two mechanical models of blood: viscous Newtonian and non-local non-Newtonian fluids. The blood flow was a Poiseuille flow through an axi-symmetric circular cylinder whose wall was rigid and permeable only to NO. Numerical simulations suggest that the concentration of NO in the endothelium is higher at higher gradients of the cerebral blood pressure. This is a very promising result since it could help understand the effects of high blood pressure on the NO concentration. Incorporating the NO production by deoxygenated red blood cells and the viscoelasticity of the endothelium in the model and performing a sensitivity analysis of the model's parameters could provide more accurate predictions of the spatio-temporal distribution of the NO concentration in brain, and thus these will be the focus of future work.

References

1. Attwell, D., Buchan, A., Charpak, S., Lauritzen, M., MacVicar, B.A., Newman, E.A.: Glial and neuronal control of brain blood flow. *Nature* **468**, 232–243 (2010). <https://doi.org/10.1038/nature09613>
2. Barbosa, R.M., Lourenco, C.F., Santos, R.M., Pomerleau, F., Huettl, P., Gehardt, G.A., Laran-jinha, J.: In vivo real-time measurement of nitric oxide in anesthetized rat brain. *Methods Enzymol.* **441**, 351–367 (2008). [https://doi.org/10.1016/S0076-6879\(08\)01220-2](https://doi.org/10.1016/S0076-6879(08)01220-2)
3. Bizjak, D.A., Brinkmann, C., Bloch, W., Grau, M.: Increase in red blood cell-nitric oxide synthase dependent nitric oxide production during red blood cell aging in health and disease: a study on age dependent changes of rheologic and enzymatic properties in red blood cells. *PLoS ONE* **10**(4) (2015). <https://doi.org/10.1371/journal.pone.0125206>
4. Bryan, N.S.: Nitric oxide enhancement strategies. *Future Sci. OA.* **1**(1), FSO48 (2015). <https://doi.org/10.4155/FSO.15.48>
5. Buerk, D.G.: Can we model nitric oxide biotransport? A survey of mathematical models for a simple diatomic molecule with surprisingly complex biological activities. *Annu. Rev. Biomed. Eng.* **3**, 109–143 (2001). <https://doi.org/10.1146/annurev.bioeng.3.1.109>
6. Buerk, D.G., Ances, B.M., Greenberg, J.H., Detre, J.A.: Temporal dynamics of brain tissue nitric oxide during functional forepaw stimulation in rats. *NeuroImage* **18**, 1–9 (2003)
7. Contestabile, A., Monti, B., Polazzi, E.: Neuronal-glia interactions define the role of nitric oxide in neural functional processes. *Curr. Neuropharmacol.* **10**(4), 303–310 (2012). <https://doi.org/10.2174/157015912804143522>
8. Drapaca, C.S.: Poiseuille flow of a non-Local non-Newtonian fluid with wall slip: a first step in modeling cerebral microaneurysms. *Fractal. Fract.* **2**(9) (2018). <https://doi.org/10.3390/fractalfract2010009>
9. Drapaca, C.S., Zhang, Z., Meng, R.: A comparison of constitutive models of blood (2018). [arXiv:1808.07977](https://arxiv.org/abs/1808.07977)
10. Forstermann, U., Sessa, W.C.: Nitric oxide synthases: regulation and function. *Eur. Heart J.* **33**, 829–837 (2012). <https://doi.org/10.1093/eurheartj/ehr304>
11. Garry, P.S., Ezra, M., Rowland, M.J., Westbrook, J., Pattinson, K.T.S.: The role of the nitric oxide pathway in brain injury and its treatment—from bench to bedside. *Exp. Neurol.* **263**, 235–243 (2015). <https://doi.org/10.1016/j.expneurol.2014.10.017>
12. Gordon, G.R.J., Mulligan, S.J., MacVicar, B.A.: Astrocyte control of the cerebrovasculature. *Glia* **55**, 1214–1221 (2007). <https://doi.org/10.1002/glia.20543>
13. Gordon, G.R., Howarth, C., MacVicar, B.A.: Bidirectional control of arteriole diameter by astrocytes. *Exp. Physiol.* **96**(4), 393–399 (2011). <https://doi.org/10.1113/expphysiol.2010.053132>

14. Hall, C.N., Garthwaite, J.: Inactivation of nitric oxide by rat cerebellar slices. *J. Physiol.* **577**(2), 549–567 (2006). <https://doi.org/10.1113/jphysiol.2006.118380>
15. Helms, C.C., Liu, X., Kim-Shapiro, D.B.: Recent insights into nitrite signaling processes in blood. *Biol. Chem.* **3**, 319–329 (2016). <https://doi.org/10.1515/hsz-2016-0263>
16. Kavdia, M., Tsoukias, N.M., Popel, A.S.: Model of nitric oxide diffusion in an arteriole: impact of hemoglobin-based blood substitute. *Am. J. Physiol. Heart Circ. Physiol.* **282**, H2245–H2253 (2002). <https://doi.org/10.1152/ajpheart.00972.2001>
17. Klabunde, R.E.: Cardiovascular physiology concepts (2017). <https://www.cvphysiology.com/Hemodynamics/H011>. Cited 11 Sept 2019
18. Ledo, A., Barbosa, R.M., Gerhardt, G.A., Cadenas, E., Laranjinha, J.: Concentration dynamics of nitric oxide in rat hippocampal subregions evoked by stimulation of the NMDA glutamate receptor. *Proc. Natl. Acad. Sci. USA* **102**(48), 17483–17488 (2005). <https://doi.org/10.1073/pnas.0503624102>
19. Long, D.S., Smith, M.L., Pries, A.R., Ley, K., Damiano, E.R.: Microviscometry reveals reduced blood viscosity and altered shear rate and shear stress profiles in microvessels after hemodilution. *Proc. Natl. Acad. Sci. USA* **101**(27), 10060–10065 (2004). <https://doi.org/10.1073/pnas.0402937101>
20. Mishra, A.: Binaural blood flow control by astrocytes: listening to synapses and the vasculature. *J. Physiol.* **595**(6), 1885–1902 (2017). <https://doi.org/10.1113/JP270979>
21. Moncada, S., Palmer, R.M.J., Higgs, E.A.: Nitric oxide: physiology, pathophysiology, and pharmacology. *Pharmacol. Rev.* **43**(2), 109–142 (1991)
22. Santos, R.M., Lourenco, C.F., Ledo, A., Barbosa, R.M., Laranjinha, J.: Nitric oxide inactivation mechanisms in the brain: role in bioenergetics and neurodegeneration. *Int. J. Cell Biol.* (2012). <https://doi.org/10.1155/2012/391914>
23. Shampine, L.F., Kierzenka, J.: A BVP solver that controls residual and error. *J. Numer. Anal. Ind. Appl. Math.* **3**(1–2), 27–41 (2008)
24. Sheps, S.G.: Pulse pressure: an indicator of heart health? (2019). <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/pulse-pressure/faq-20058189>. Cited 11 Sept 2019
25. Sriram, K., Laughlin, J.G., Rangamani, P., Tartakovsky, D.M.: Shear-induced nitric oxide production by endothelial cells. *Biophys. J.* **111**, 208–221 (2016). <https://doi.org/10.1016/j.bpj.2016.05.034>
26. Vaughn, M.W., Kuo, L., Liao, J.C.: Effective diffusion distance of nitric oxide in the microcirculation. *Vasc. Physiol.* **274**(5), H1705–H1714 (1998). <https://doi.org/10.1152/ajpheart.1998.274.5.H1705>
27. Vaughn, M.W., Kuo, L., Liao, J.C.: Estimation of nitric oxide production and reaction rates in tissue by use of a mathematical model. *Am. J. Physiol.* **274**(6), H2163–H2176 (1998). <https://doi.org/10.1152/ajpheart.1998.274.6.H2163>

Automate Obstructive Sleep Apnea Diagnosis Using Convolutional Neural Networks



Longlong Feng and Xu Wang

Abstract Identifying sleep problem severity from overnight polysomnography (PSG) recordings plays an important role in diagnosing and treating sleep disorders such as the Obstructive Sleep Apnea (OSA). This analysis traditionally is done by specialists manually through visual inspections, which can be tedious, time-consuming, and is prone to subjective errors. One of the solutions is to use Convolutional Neural Networks (CNN) where the convolutional and pooling layers behave as feature extractors and some fully-connected (FCN) layers are used for making final predictions for the OSA severity. In this paper, a CNN architecture with 1D convolutional and FCN layers for classification is presented. The PSG data for this project are from the Cleveland Children's Sleep and Health Study database and classification results confirm the effectiveness of the proposed CNN method. The proposed 1D CNN model achieves excellent classification results without manually preprocessing PSG signals such as feature extraction and feature reduction.

Keywords Deep learning · Convolutional neural network · Polysomnography · Obstructive sleep apnea

1 Introduction and Background

When we sleep, our muscles relax. For the Obstructive Sleep Apnea (OSA) patients, the muscles in the back of throat can relax too much and collapse the airway, and lead to breathing difficulty. OSA presents with abnormal oxygenation, ventilation

L. Feng (✉)

Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: feng0290@mylaurier.ca

X. Wang

Department of Mathematics & MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: xwang@wlu.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_48

521

and sleep pattern. The prevalence of OSA has been reported to be between 1% to 5% [1]. Children at risk need timely investigation and treatment.

The gold standard for diagnosing sleep disorders is polysomnography (PSG), which generates extensive data about biophysical changes during sleep. Studies of PSG assist doctors to diagnose sleep disorders and provide the baseline for an appropriate follow up. A clinical sleep study design based on PSG is to acquire several biological signals while patients are sleeping, These signals typically include electroencephalography (EEG) for monitoring brain activity, electromyogram (EMG) to measure muscle activity and Electrocardiography (ECG) for the electrical activity of heart over a period of sleep [2].

In recent decades, various alternative methods have been proposed to minimize the number of biosignals required to detect and classify the OSA. These studies include traditional machine learning methods such as Support Vector Machine and linear discriminant analysis on signals such as ECG [3], respiratory signals [4], a combination of extracted features and shallow neural network on heart rate variability and ECG derived respiration signal [5]. These studies focused on extracting time domain, frequency domain, and other nonlinear features from physiological signals and applying some feature selection techniques to reduce the number of dimensions comprising the feature space. However, this process can be labour-intensive, requires domain knowledge, and is particularly limited and costly for high-dimensional data. In addition, feature extraction is difficult for traditional machine learning techniques as the number of features increase dramatically.

Deep learning framework has proved its modeling ability in different PSG channels. McCloskey et al. employed a 2D-CNN model on spectrograms of nasal airflow signal, and their model achieved an average accuracy of 77.6% on three severity levels [6]. Another more outstanding application of deep learning model came from the work of Cheng et al. in which researchers used a four layered Long Short Term Memory (LSTM) model on the RR-ECG signal and achieved an average accuracy of 97.80% on the detection of OSA [7].

Though recurrent model (e.g., RNN, LSTM) can process time-series data and make sequential predictions, CNN can be trained to recognize the *same* patterns (severity levels) on different subfields within fixed time windows. CNN saves time from manual scoring in the laboratory environment and makes the pre-screening stage easier in contrast to traditional methods. Moreover, in order to increase the model generalization ability, we tried to explore 1D-CNN models with different length of segmentations in EEG, ECG, EMG and respiratory channels. We focused on the model structure and utilized the fine-tuned model for pediatric OSA prediction in our study.

The rest of this paper is organized as follows. Section 2 explains the data processing in detail. Section 3 displays the structure of the proposed 1D-CNN model. Evaluation and experimental results are presented in Sect. 4. Finally, Sect. 5 draws discussion and conclusion of the research.

2 Cleveland Children’s Sleep and Health Study Database

The data are retrieved from the National Sleep Research Resource (NSRR), which is a new National Heart, Lung, and Blood Institute resource designed to provide big data resources to the sleep research community. The PSG data are available from Cleveland Children’s Sleep and Health Study (CCSHS) database. Each anonymous record includes a summary result of a 12-hour overnight sleep study (awake and sleep stages) including annotation files with scored events and PSG signals and being formatted as the European Data Format (EDF).

The following channels are selected for the 1D CNN Modeling: 4 EEG channels (*C3/C4* and *A1/A2*), 3 EMG channels (*EMG1*, *EMG2*, *EMG3*), 2 ECG channels (*ECG1* and *ECG2*), and 3 respiratory channels including airflow, thoracic and abdominal breathing.

2.1 Individual Labeling

To define the target variable for this classification problem, each participant needs one label based on the OSA severity level. The Obstructive Apnea Hypopnea Index (*oahi3*) is used to indicate the severity of sleep apnea. It is represented by the number of apnea and hypopnea events per hour of sleep. It combines AHI and oxygen desaturation to give an overall sleep apnea severity score that evaluates both the number of sleep disruptions and the degree of oxygen desaturation (low oxygen level in the blood). The values of *oahi3* are used as the thresholds for grouping the participants. The number of participants with different severity levels are shown in Table 1.

The dataset has an imbalanced response variable (362 normal/139 minor/8 moderate/8 severe). Those minority classes (moderate and severe) are our most interest. We tried to train classifier to learn more from moderate and severe level data. Under-sampling method was applied during the data pre-processing stage, i.e., we randomly selected an equal number of samples (i.e., 8 participants) from each of the normal and minor groups. Overall, there are 32 participants in the final study data set. In this project, we conduct data pre-processing and CNN modeling on the data in EDF format which have a total size of 13 GB.

Table 1 Grouping participants using *oahi3* values

Obstructive Apnea Hypopnea Index	Level of severity	Number of participants
$0 < \text{oahi3} \leq 1$	NL (Normal)	362
$1 < \text{oahi3} \leq 5$	MIN (Minor)	139
$5 < \text{oahi3} \leq 10$	MOD (Moderate)	8
$10 < \text{oahi3}$	SV (Severe)	8

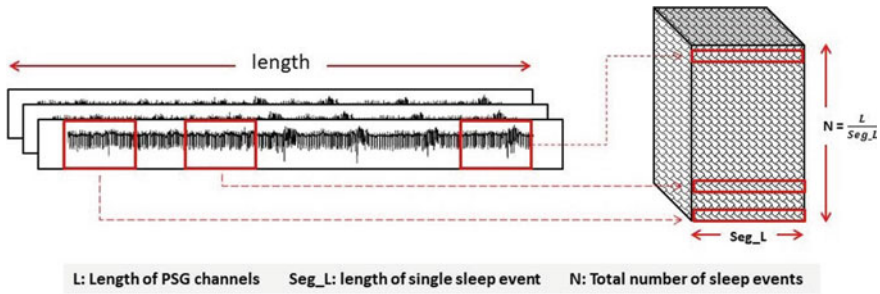


Fig. 1 Demonstration of channel division

2.2 Data Preprocessing

This experiment focuses on the sleep data. The beginning and ending *awake* signals could be treated as noise and need to be removed. Secondly, the deep learning algorithms tend to be difficult to train when the length of time series is very long. Figure 1 presents a segmentation strategy, i.e., dividing the time series into smaller chunks.

Each segment was labeled as the same severity level as the participant. In other words, the segments would inherit the severity label from the participant they belong to. With a starting length of L time steps, one channel is divided into blocks of sequence Seq_L yielding about L / Seq_L of new events (or rows) of shorter length (N).

The PSG data were segmented into 1-min long events. For the ECG channel (frequency of 256) a 1-min event has a length of 15360 (256×60) data points. An individual has a 8.24-h ECG channel, which would have 1D time series data with length of 7595520. After segmentation, the long series data turned into a tensor with dimension 494×15360 , which indicates 494 events (a length of 15360 for each). Since we have 32 selected participants and 2 ECG channels for each participant, the input tensor has the dimension of $15824 (N) \times 15360 (Seq_L) \times 2$ (channels).

With the data segmentation, the length of each time-series is shorter and will be helpful in model training; and the number of data points has increased by a factor of L / Seq_L (number of instances or rows) providing a larger data set to train on.

Since different channels (e.g., ECG, EMG) were measured in different amplitudes, therefore, the last step of data processing is to normalize the PSG data with zero mean and unit standard deviation.

3 1D-CNN Architecture

The convolutional layer and max-pooling layer play the key roles in the CNNs feature extraction mechanism. The output of convolutional layer of the ℓ th layer can be calculated as in Formula 1:

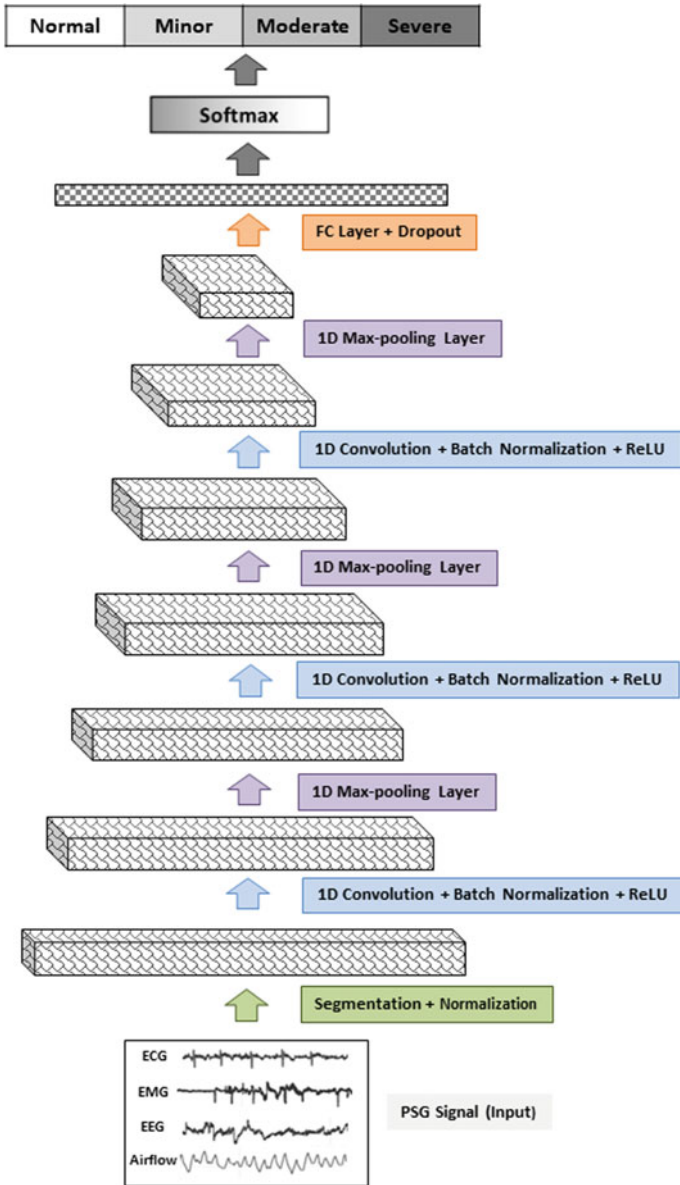


Fig. 2 The proposed 1D-CNN architecture

$$C_k^{(\ell)} = ReLU\left(\sum_c W_k^{(\ell),c} * X^{(\ell-1),c} + B_k^\ell\right), \quad (1)$$

where k represents the filter number, c denotes the channel number of the input $X^{\ell-1}$, $W_k^{(\ell),c}$ is the k th convolutional filter to the c th channel, and B_k^ℓ is the bias to the k th filter, and $*$ is the dot product operation.

The max-pooling layer is a sub-sampling function selecting the maximum value within a fixed size filter. After the convolution-pooling blocks, one fully connected layer of neurons which have full connections to all activations in the previous layer, as in the regular Neural Networks. At the end of the convolutional layers, the data were flattened and passed onto the Dropout layer before the softmax classifier.

Figure 2 shows the structure of the 1D CNN model proposed in this project. It contains 3 convolutional and 3 max-pooling layers. We focused our efforts on the CNN building and began the investigation of the CNN method initially by performing a grid search of several hyperparameters.

For each participant, his or her PSG data were served for either training or test data, not for both. We implemented a two-level stratified random sampling. In details, there were 2 splitting steps among 32 participants: firstly, 8 were randomly selected as test participants (i.e., 2 participants were randomly selected for each severity level); secondly, the remaining 24 participants were split into two groups: 18 participants for training set and 6 participants for validation set. The tensorflow graph was fed with batches of the training data and the hyperparameters were tuned on a validation set. Finally the trained model was evaluated on the test set.

The CNN model was trained in a fully supervised manner, and the gradients were back-propagated from the softmax layer to the convolutional layers. The network parameters were optimized by minimizing the cross-entropy loss function based on the gradient descent with the Adam updating rule and a learning rate of 0.0001.

Table 2 presents the final values of parameters within each layer. Dropout rate of 0.5 was used as it is the general setting for CNN models. Model classification performance is evaluated by using the following metrics: classification accuracy, cross-entropy loss, precision, recall and F1-score. While accuracy and loss can be used for evaluating the overall performance, some other metrics can be used to measure the performance of specific class.

4 Results and Analysis

Figure 3 shows the learning curve on training and validation phases. Accuracy and loss were obtained with various number of iterations. The accuracy increases as the number of iteration increases, and the loss decreases at the same time. The accuracy and the loss reach stable values after iterative learning on both phases.

For ECG, we can observe the stable accuracy and loss values after 1000 iterations (Training acc: 0.9987, loss: 0.0114; Validation acc: 0.9916, loss: 0.0289). For EEG, the accuracy and the loss start to converge to a value after 2500 iterations (Train-

Table 2 CNN model structure with optimal parameters

CNN layer	# of filters	Filter size	Stride	Padding	Activation function
Conv 1	46	10	2	No	Relu
Pooling 1	–	10	2	No	–
Conv 2	92	10	2	No	Relu
Pooling 2	–	10	2	No	–
Conv 3	184	20	2	No	Relu
Pooling 3	–	20	5	No	–

ing acc: 0.9718, loss: 0.0945; Validation acc: 0.9447, loss: 0.1985). For EMG, the accuracy and the loss become stable after 4000 iterations (Training acc: 0.9999, loss: 0.0013; Validation acc: 0.9707, loss: 0.1131). However, there are a large number of big fluctuations before the convergence during the learning process. This means some portion of the randomness: (1) The Dropout method could cause the network to keep only some portion of neurons (weights) on each iteration. Sometimes those neurons do not fit the current batch well, and this may cause large fluctuations; (2) There is randomness in initialization and data sampling for SGD in back-propagation.

For Respiratory, we can see the train and validation accuracy begin to stay steady with similar values indicating slight overfitting in the classification (Training acc: 0.9854, loss: 0.0378; Validation acc: 0.9180, loss: 0.2945).

The evaluation metrics and confusion matrices for all channels with training and test data are presented in Tables 3 and 4 respectively. The results from Table 4 are summarized in Table 3. It can be observed from Table 3 that, for the test data, the CNN model can achieve 98.97% for ECG, 94.63% for EEG, 95.81% for EMG, and 91.99% for Respiratory; We can also verify the training curves from Fig.3 by checking the training accuracy score from Table 3 and the classified results from Table 4. Furthermore, the precision, recall and F1-score for each class are collected in Table 3.

For ECG, the model can achieve a value of >99% for all three metrics for all classes on the training data and >97% for the test data; For EEG, the model achieves a >96% score for training data, and >91% for the test data.

For EMG, the scores of 1.0000 are obtained in the training phase on all classes, which means the perfect classification for the training data during the learning process, while the scores of >93.29% are obtained from the test data.

Similarly, for Respiratory, CNN achieves scores of >98% for the training and slightly lower scores, which are over >88.99% for the test data. The reason why there exists the gap between training and test scores can be that the respiratory signal sensors is different from ECG, EEG and EMG. In this case, the signal in the respiratory system may not be sensitive enough to detect small changes when OSA happens. Table 4 displays the classification details on the training and test data.

Table 3 The CNN evaluation metrics

Channels(#)	Dataset	Accuracy	Loss	Class	Precision	Recall	F1-score
ECG (2)	Training	0.9987	0.0114	NL	0.9997	0.9997	0.9994
				MIN	0.9980	0.998	0.9980
				MOD	0.9982	0.9982	0.9982
				SV	0.9988	0.9994	0.9991
	Test	0.9897	0.0289	NL	0.9862	0.9921	0.9891
				MIN	0.9990	0.9773	0.9880
				MOD	0.9894	0.9961	0.9927
				SV	0.9843	0.9940	0.9891
EEG (4)	Training	0.9718	0.0945	NL	0.9753	0.9741	0.9747
				MIN	0.9784	0.9820	0.9802
				MOD	0.9721	0.9684	0.9703
				SV	0.9609	0.9621	0.9615
	Test	0.9463	0.1985	NL	0.9394	0.9587	0.9490
				MIN	0.9415	0.9741	0.9575
				MOD	0.9682	0.9166	0.9417
				SV	0.9373	0.9354	0.9363
EMG (3)	Training	0.9999	0.0013	NL	1.0000	1.0000	1.0000
				MIN	1.0000	0.9997	0.9999
				MOD	0.9997	1.0000	0.9999
				SV	1.0000	1.0000	1.0000
	Test	0.9581	0.1132	NL	0.9518	0.9312	0.9414
				MIN	0.9660	0.9601	0.9631
				MOD	0.9823	0.9712	0.9767
				SV	0.9329	0.9696	0.9509
Respiratory (3)	Training	0.9854	0.0378	NL	0.9857	0.9857	0.9857
				MIN	0.9834	0.9849	0.9842
				MOD	0.9895	0.9880	0.9888
				SV	0.9828	0.9828	0.9828
	Test	0.9199	0.2945	NL	0.9147	0.9147	0.9147
				MIN	0.9447	0.9053	0.9246
				MOD	0.9323	0.9194	0.9258
				SV	0.8899	0.9408	0.9147

Note NL (Normal), MIN (Minor), MOD (Moderate), SV (Severe)

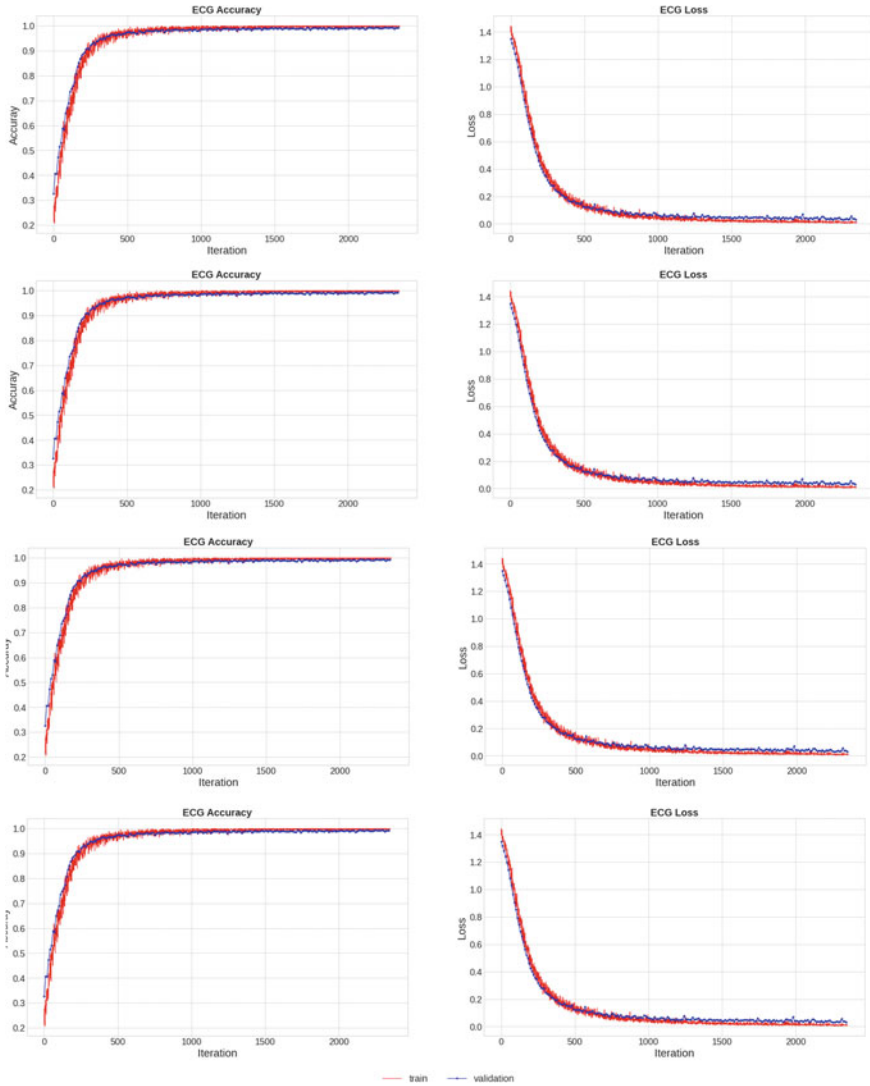


Fig. 3 Accuracy and loss of the proposed CNN model for OSA detection

5 Conclusion and Discussion

Firstly, with the correct hyper-parameter setup, our 1D-CNN model can successfully extract the temporal features from the PSG data and achieve high performance in OSA detection for different channels; secondly, our well trained CNN model can be an efficient tool for clinicians to identify OSA severity without manually going through tons of PSG data. Furthermore, our CNN models can replace the traditional data

Table 4 Confusion matrices from the CNN model on training and test data

		ECG training				ECG test			
True	Predict								
	NL	MIN	MOD	SV	NL	MIN	MOD	SV	
NL	3321	0	1	2	1000	0	3	5	
MIN	0	3511	5	2	10	1032	6	8	
MOD	1	5	3378	0	1	0	1022	3	
SV	0	2	0	3340	3	1	2	1000	
		ECG training				ECG test			
True	Predict								
	NL	MIN	MOD	SV	NL	MIN	MOD	SV	
NL	3239	14	13	59	976	12	4	26	
MIN	9	3445	34	20	7	1014	11	9	
MOD	15	40	3279	52	28	30	945	28	
SV	58	22	47	3222	28	21	16	941	
		ECG training				EMG test			
True	Predict								
	NL	MIN	MOD	SV	NL	MIN	MOD	SV	
NL	3546	0	0	0	731	14	9	31	
MIN	0	3745	1	0	11	795	5	17	
MOD	0	0	3601	0	9	7	775	7	
SV	0	0	0	3571	17	7	0	765	
		Respiratory training				Respiratory test			
True	Predict								
	NL	MIN	MOD	SV	NL	MIN	MOD	SV	
NL	3714	18	14	22	461	10	14	19	
MIN	16	3912	13	31	10	478	14	26	
MOD	17	17	3786	12	20	7	468	14	
SV	21	31	13	3723	13	11	6	477	

Note NL (Normal), MIN (Minor), MOD (Moderate), SV (Severe)

processing such as signal extraction and transforming, which can be time-consuming and labour-intensive.

There are some limitations of our work. Firstly, only a small sample of 32 subjects was investigated in this study. Secondly, we used ECG, EEG, EMG and Respiratory channels to build CNN models separately, so there was no cross-checking between different channels. Lastly, our CNN model is slow to be trained without GPU. The well-trained models require a big data set and the fine-tuned hyperparameters in the training step.

The future work can aim at feeding the four single CNN models into an ensemble-like model to making a prediction. There are other possible architectures that would be of great interest for this problem. One of most popular deep learning architectures

that models sequence and time-series data is the long-short-term memory (LSTM) cells within recurrent neural networks (RNN).

Acknowledgements We are so grateful that National Sleep Research Resource (NSRR) allows us to use the PSG data from Cleveland Children’s Sleep and Health Study. The project is supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Dehlink, E., Tan, H.-L.: Update on paediatric obstructive sleep apnoea. *J. Thorac. Dis.* **8**(2) (2016)
2. Moridani, M.K., Heydar, M., Jabbari Behnam, S.S.: A reliable algorithm based on combination of EMG, ECG and EEG signals for sleep apnea detection: (a reliable algorithm for sleep apnea detection). In: 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEL), pp. 256–262 (2019)
3. Almuhammadi, W.S., Aboalayon, K.A.I., Faezipour, M.: Efficient obstructive sleep apnea classification based on EEG signals. In: 2015 Long Island Systems, Applications and Technology, pp. 1–6 (2015)
4. Varon, C., Caicedo, A., Testelmans, D., Buyse, B., Van Huffel, S.: A novel algorithm for the automatic detection of sleep apnea from single-lead ECG. *IEEE Trans. Biomed. Eng.* **62**(9), 2269–2278 (2015)
5. Tripathy, R.K.: Application of intrinsic band function technique for automated detection of sleep apnea using HRV and EDR signals. *Biocybern. Biomed. Eng.* **38**(1), 136–144 (2018)
6. McCloskey, S., Haidar, R., Koprinska, I., Jeffries, B.: Detecting Hypopnea and Obstructive Apnea Events Using Convolutional Neural Networks on Wavelet Spectrograms of Nasal Airflow, pp. 361–372 (2018)
7. Cheng, M., Sori, W., Jiang, F., Khan, A., Liu, S.: Recurrent neural network based classification of ECG signal features for obstruction of sleep apnea detection, pp. 199–202 (2017)

Numerical Analysis of Nanowire Resonators for Ultra-high Resolution Mass Sensing in Biomedical Applications



Rosa Fallahpour and Roderick Melnik

Abstract Nanowire resonators have fascinated researchers as a promising group of devices for accurate detection of tiny objects such as atoms, molecules, viruses, bacteria, and different types of bio-objects. In this paper, we present a numerical solution to the newly developed mathematical model of the nanowire resonator, considering such important characteristics as temperature variations, as well as the electromagnetic fields, added mass, surface and nonlocal effects. The mathematical model is based on the nonlocal Euler-Bernoulli beam theory. The developed model is solved by using the Finite Difference Method (FDM). As a result of this solution, the frequency response of the nanowire resonator has been obtained. Then, based on the developed numerical solution, a parametric study has been carried out to investigate the effects of different parameters on the vibration of nanowire resonators. Finally, the importance of nonlinearity in the modelling of such resonators at the nanoscale has been highlighted.

Keywords Nanobiomedicine · Mathematical modelling · Nonlocal nonlinear problems · Euler-bernoulli beam theory · Nanowire resonators · Bio-object detection

1 Introduction

The extraordinary and unique mechanical properties of nanostructures have made them an excellent candidate for mass sensing applications. Different types of nanores-

R. Fallahpour (✉) · R. Melnik

The MS2Discovery Interdisciplinary Research Institute, M2NeT Laboratory, Wilfrid Laurier University, Waterloo, ON N2L3C5, Canada
e-mail: rosa.fp68@gmail.com

R. Melnik
e-mail: rmelnik@wlu.ca

R. Melnik
BCAM-Basque Center for Applied Mathematics, Alda. Mazarredo, 48009 Bilbao, Spain

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_49

533

onators including nanowires, quantum dots, nanotubes and graphene sheets have been studied by researchers from different fields to be used for tiny object detection. In particular, the ultra-high frequency of these structures attracted attention of researchers in the area of bio-sensing to implement them for the detection of tiny bio-objects. As a result, plenty of theoretical and practical approaches have been proposed for the detection of tiny bio-particles. By using both theoretical and experimental approaches, researchers have shown a strong potential of nanoresonators in the detection of tiny objects even in the scale of zeptogram (zg) [1–3].

Nanoresonators, specifically semi-conducting nanowires, have shown very unique reproducible and tunable conducting properties, which provide a basis for strong sensing approaches in medical applications [4]. This high resolution of sensing allows detection of tiny bio-objects such as DNA, RNA, proteins, viruses, bacteria and very small chemical atoms. Analysis of temperature variations is one of the well-addressed parameters, but in order to analyze other significant parameters, the development of novel models is needed to provide a better understanding of nanoresonator's sensing resolution. Thus, vibration characterization and parametric sensitivity analysis of nano-mechanical resonators for sensing applications are crucial, notably in biomedicine. It should be noted that when we refer to nanosensors, we deal with a resonator with dimensions in the order of nanometer, which has sensitivity in the nanoscale range, and its interaction distance with the object being detected in nanometer size. That is why a small perturbation with different sources of excitation such as temperature, electromagnetic field, nonlinearity due to large oscillations of the nanoresonators or their substrates should be taken into account for an adequate mathematical modelling of these devices. Accordingly, vibrations of nanoresonators including nanobeams, quantum dots, nanotubes, nanowires, graphene sheets, and nanoplates have been receiving an increased attention in the interdisciplinary community of researchers, including those working in the areas of applied mechanics and mathematics, structural analysis and vibrations. A number of works have been published so far to investigate the vibrations of nanoresonators [5–8].

An analysis of the state-of-the-art in this field shows that there is a lack of modelling results for nanowire resonators in mass detection applications that take into account different critical parameters such as the electromagnetic fields, piezoelectric potential, nonlinearity, external excitations and thermal variations. This shortcoming of current knowledge in this area has prompted us to work on the development, as well as on mathematical and numerical analysis, of a novel continuum model for nanowire resonators.

In this article, we briefly describe our developed mathematical model for the vibrations of nanowire resonators. Our proposed model is based on the nonlocal Euler-Bernoulli beam theory and includes the terms related to the added mass, temperature variations, electromagnetic fields, large oscillations, and piezoelectric effect. Then, a finite difference scheme is developed to obtain the natural frequency of the nanowire resonator. Finally, a parametric sensitivity analysis is presented to show the effect of different parameters on the frequency behavior of nanowire resonators with an added mass, and the importance of nonlinearity is highlighted.

2 Mathematical Modelling

Utilizing the nonlocal Euler Bernoulli beam theory and incorporating different effects including surface, electromagnetic field, thermal variations, large oscillations, added mass and nonlinear foundation, the following nonlinear partial differential equation is developed for the vibrations of nanowires [9]:

$$(EI)_{eff} \frac{\partial^4 w}{\partial x^4} + \left(1 - \Gamma \frac{\partial^2}{\partial x^2}\right) \Psi = 0, \tag{1}$$

where

$$\begin{aligned} \Psi = & (\rho A)_{eff} \frac{\partial^2 w(x, t)}{\partial t^2} + m_p \delta(x - x_p) \frac{\partial^2 w(x, t)}{\partial t^2} + \\ & \mu \frac{\partial w(x, t)}{\partial t} + k_1 w(x, t) - 2b\tau_0 \frac{\partial^2 w}{\partial x^2} + k_3 w^3(x, t) - F(x, t) - \zeta_m A H_x^2 \frac{\partial^2 w}{\partial x^2} + \\ & 2V_e b e_{31} \frac{\partial^2 w}{\partial x^2} - \left(\frac{(EA)_{eff}}{2L} \int_0^L \left(\frac{\partial w}{\partial x}\right)^2 dx - N_\theta\right) \frac{\partial^2 w}{\partial x^2}. \end{aligned} \tag{2}$$

The definition of other terms presented in Eqs. (1–2) can be found in Ref. [9]. In order to solve the above partial differential equation, we develop a finite difference approximation. In the next section, we briefly describe the numerical approach applied to the solution of this problem. The considered boundary are described in Eq. (3):

$$W(0, t) = 0, \frac{\partial W}{\partial x}(0, t) = 0, W(L, t) = 0, \frac{\partial W}{\partial x}(L, t) = 0, \tag{3}$$

and the following general form of initial conditions are assumed:

$$W(x, t = 0) = W_0, \frac{\partial W}{\partial t}(x, t = 0) = \bar{W}_0, \tag{4}$$

where W_0 and \bar{W}_0 are given functions. Motivated by the applications of interest here, the model (1)–(4) is simplified in the next section. Assuming periodicity in time, we propose a solution procedure where the function W will be analyzed with respect to frequency rather than time, moving our consideration to the frequency domain.

3 Solution Procedure

In this section we concisely illustrate the FDM in the context of our problem, and then move to the implementation of this method for the nanowire resonator.

3.1 FDM

Finite difference methods are a generic class of numerical methods, which are used for solving differential equations by approximating them with difference equations, where finite differences approximate the derivatives. FDMs require a discretization of the computational domain. The domain is partitioned in both space (x) and time (t), and approximations of the solution are computed at points of the grid, resulted from the domain discretization. Based on the FDM, the discretized equations for the first, second, third and fourth derivatives with respect to x are as follows [10–12]:

$$\frac{\partial w}{\partial x} \approx \frac{w_{i+1} - w_{i-1}}{2\Delta x}, \quad \frac{\partial^2 w}{\partial x^2} \approx \frac{w_{i+1} - 2w_i + w_{i-1}}{(\Delta x)^2}, \quad (5)$$

$$\frac{\partial^3 w}{\partial x^3} \approx \frac{w_{i+3} - 3w_{i+2} + 3w_{i+1} - w_i}{(\Delta x)^3}, \quad (6)$$

$$\frac{\partial^4 w}{\partial x^4} \approx \frac{w_{i-2} - 4w_{i-1} + 6w_i - 4w_{i+1} + w_{i+2}}{(\Delta x)^4}, \quad (7)$$

where

$$\Delta x = \frac{\text{Length of } X}{\text{Number of Steps in } X}. \quad (8)$$

The accuracy of approximations (5) and (7) at the grid point x_i , is of the second order, and approximation (6) is of the first order, with respect to (Δx) . In the next sub-section, we apply the FDM to our developed governing equation, Eq. (1), to allow a numerical analysis of the frequency of nanowire resonators.

3.2 Implementation of the FDM for the Nanowire Resonator

In this part, we apply the FDM to the governing equation (Eq. (1)) of nanowire resonators. In order to use the FDM to analyze the developed model, we assume that the displacement of the nanowire resonator can be given in the following form [10]:

$$w(x, t) = w(x)e^{i\omega t}, \quad (9)$$

where ω is the frequency of the nanowire resonator. We first consider the linear part of the developed Eq. (1) [9]. Substituting Eq. (9) into the linear part of Eq. (1) results in:

$$(P) \frac{\partial^4 w(x)}{\partial x^4} + (Q) \frac{\partial^2 w(x)}{\partial x^2} + k_1 w(x) = \bar{M} \omega^2 w(x). \quad (10)$$

By substituting the approximate derivatives, Eq. (5) and Eq. (7), into Eq. (10), the following form is obtained:

$$G_1(w_{i-2} - 4w_{i-1} + 6w_i - 4w_{i+1} + w_{i+2}) + G_2(w_{i+1} - 2w_i + w_{i-1}) + k_1 w_i = \omega^2 [-G_3 w_i - G_4(w_{i+1} - 2w_i + w_{i-1})], \quad (11)$$

where

$$G_1 = \frac{\left[(EI)_{eff} + 2\Gamma b\tau_0 + \Gamma \zeta AH_x^2 + 2\Gamma vbe_3 1 + \Gamma \frac{(EA)_{eff}}{2L} N_\theta \right]}{(\Delta x)^4}, \quad (12)$$

$$G_2 = \frac{\left[-2b\tau_0 - \zeta_m AH_x^2 - 2vbe_3 1 - \frac{(EA)_{eff}}{2L} N_\theta - \Gamma k_1 \right]}{(\Delta x)^2}, \quad (13)$$

$$G_3 = \rho A + m_p, \quad G_4 = \Gamma \frac{m_p + \rho A}{\Delta x^2}, \quad (14)$$

and

$$\bar{M} = [-G_3 w_i - G_4(w_{i+1} - 2w_i + w_{i-1})]. \quad (15)$$

Using Eq. (8) for $i = 1, \dots, N$, we can represent Δx as below:

$$\Delta x = \frac{L}{(N-1)}, \quad (16)$$

where L is the length of the nanowire resonator. By considering clamped-clamped boundary conditions for the nanoresonator, we will have the following equations for both ends of the nanowire:

$$\text{at } x = 0 : w_1 = 0, \quad \& \quad \frac{w_2 - w_0}{2\Delta x} = 0, \quad (17)$$

and

$$\text{at } x = N : w_N = 0, \quad \& \quad \frac{w_{N+1} - w_{N-1}}{2\Delta x} = 0. \quad (18)$$

Based on Eqs. (17) and (18), we obtain the following relations:

$$w_0 = w_2, \quad w_{N+1} = w_{N-1}. \quad (19)$$

It should be mentioned that w_0 and w_{N+1} are fictitious values which can be eliminated in our governing equation by using Eq. (19). Now, in order to solve Eq. (10) using the FDM, we substitute $i = 2, \dots, N-1$ into Eq. (11), which results in a system of equations as follows:

$$\begin{bmatrix} A_1 & B_1 & C_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ A_2 & B_2 & C_2 & D_2 & \dots & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & A_{N-3} & B_{N-3} & C_{N-3} & D_{N-3} \\ 0 & 0 & 0 & 0 & \dots & 0 & A_{N-2} & B_{N-2} & C_{N-2} \end{bmatrix} \begin{bmatrix} w_2 \\ w_3 \\ \cdot \\ \cdot \\ \cdot \\ w_{N-2} \\ w_{N-1} \end{bmatrix} = \omega^2 \quad (20)$$

$$\begin{bmatrix} G_3 - 2G_4 & G_4 & \dots & 0 & 0 \\ G_4 & G_3 - 2G_4 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & G_3 - 2G_4 & G_4 \\ 0 & 0 & \dots & G_4 & G_3 - 2G_4 \end{bmatrix} \begin{bmatrix} w_2 \\ w_3 \\ \cdot \\ \cdot \\ \cdot \\ w_{N-2} \\ w_{N-1} \end{bmatrix},$$

where

$$A_1 = 7G_1 - 2G_2 + k_1, \quad B_1 = -4G_1 + G_2, \quad C_1 = G_1, \quad (21)$$

$$A_2 = -4G_1 + G_2, \quad B_2 = 4G_1 - 2G_2 + k_1, \quad C_2 = -4G_1 + G_2, \quad D_2 = G_1, \quad (22)$$

$$\begin{aligned} A_{N-3} &= G_1, & B_{N-3} &= -4G_1 + G_2, \\ C_{N-3} &= 6G_1 - 2G_2 + k_1, & D_{N-3} &= -4G_1 + G_2, \end{aligned} \quad (23)$$

and

$$A_{N-2} = G_1, \quad B_{N-2} = -4G_1 + G_2, \quad C_{N-2} = 7G_1 - 2G_2 + k_1. \quad (24)$$

Hence, Eq. (20) can be rewritten in the following form:

$$(-[\bar{M}]\omega^2 + [K])\{w\} = 0, \quad (25)$$

where $w = \{w_2, w_3, \dots, w_{N-2}, w_{N-1}\}^T$, \bar{M} and K are the mass and stiffness matrices, respectively. \bar{M} is defined by the following matrix:

$$\bar{M} = \begin{bmatrix} G_3 - 2G_4 & G_4 & \dots & 0 & 0 \\ G_4 & G_3 - 2G_4 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & G_3 - 2G_4 & G_4 \\ 0 & 0 & \dots & G_4 & G_3 - 2G_4 \end{bmatrix}. \quad (26)$$

K represents the stiffness matrix, and it can be found as follows:

$$K = \begin{bmatrix} A_1 & B_1 & C_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ A_2 & B_2 & C_2 & D_2 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & A_{N-3} & B_{N-3} & C_{N-3} & D_{N-3} \\ 0 & 0 & 0 & 0 & \dots & 0 & A_{N-2} & B_{N-2} & C_{N-2} \end{bmatrix}. \tag{27}$$

In order to obtain the linear frequency of the vibrations of the nanowire resonator, we need to find the solution of Eq. (25). A non-trivial solution of Eq. (25) can be obtained when the determinant of coefficient matrix equals to zero:

$$\left| [-\bar{M}]\omega^2 + [K] \right| = 0. \tag{28}$$

Based on the algorithm described above, stiffness and mass matrices can be calculated numerically. These calculated matrices are used in Eq. (28) to obtain the linear natural frequency in Eq. (25). It should be noted that the size of mass and stiffness matrices depends on the number of nodes N. For the nonlinear part of the governing equation, Eq. (1), we have:

$$NL := -\frac{EA_{eff}}{2L} \left[\int_0^L \left(\frac{\partial w(x,t)}{\partial x} \right)^2 dx \right] \frac{\partial^2 w(x,t)}{\partial x^2} + k_3 w^3(x,t) \tag{29}$$

$$-\Gamma k_3 \frac{\partial^2}{\partial x^2} [w^3(x,t)] + \Gamma \frac{EA_{eff}}{2L} \left[\int_0^L \left(\frac{\partial w(x,t)}{\partial x} \right)^2 dx \right] \frac{\partial^2 w(x,t)}{\partial x^2}.$$

The integral term in our governing equation can be approximated by the following relationship:

$$\int_0^L \left(\frac{\partial w(x,t)}{\partial x} \right)^2 dx \approx \frac{L}{2} \left[\left(\frac{\partial w(x,t)}{\partial x} \right)^2 \Big|_{i=1} + \left(\frac{\partial w(x,t)}{\partial x} \right)^2 \Big|_{i=N} \right]. \tag{30}$$

Based on the defined boundary conditions in Eqs. (17–18) and the nonlinear terms presented by Eq. (29), we have the following relation to obtain the nonlinear frequency of the considered nanowire resonator:

$$(-[\bar{M}]\omega^2 + [K] + [K_{NL}])\{w\} = 0. \tag{31}$$

To find the nonlinear natural frequency, we first need to solve the linear equation to obtain the eigenvalues and eigenvectors. It should be noted that eigenvectors and eigenvalues represent mode shapes and the linear frequencies of vibrations, respectively. Basically, these two values are used in an iterative process to obtain the nonlinear natural frequencies. Then, we utilize the obtained solution as an initial approximation to the nonlinear equation defined by Eq. (31). By substituting the derived eigenvalues and eigenvectors into Eq. (31), and also coupling the linear

and nonlinear stiffness matrices with the mass matrix, the nonlinear frequency and mode shape can be calculated [13]. Then, implementing the iteration method, the nonlinear frequency is recalculated in order to find an approximate frequency, when the iterations converge with pre-defined accuracy. In the next section, we discuss the results obtained based on the developed numerical approach.

4 Results and Discussion

In this section, a parametric sensitivity analysis is carried out by using the numerical solution obtained with the methodology described in the previous section, for the vibration of the nanowire resonator. All figures in this section are obtained based on Eq. (28) using parameters defined in Ref. [9] with clamped-clamped boundary conditions given by Eqs. (17) and (18). We have investigated the sensitivity of dimensionless frequency, \bar{f}_n , obtained by numerical simulation presented in Sect. 3 with respect to variations in temperature, piezoelectric voltage, nonlocal parameter, and the added mass. The dimensionless frequency, \bar{f}_n , is defined by using the following equation:

$$\bar{f}_n = \frac{\omega}{\omega_0}, \quad (32)$$

where both ω and ω_0 can be obtained by using Eq. (28). The constant ω_0 is the frequency of the nanoresonator without considering the effect of added mass. Figure 1a shows the effect of temperature on the frequency behavior of silicon nanowire (SiNW) resonator using Eq. (32). As this figure shows, increasing the temperature reduces the frequency value of the nanowire resonator. A linear relation is observed between the temperature rise and the frequency reduction of the nanowire resonator. Considering the developed continuum model in our analysis, the main reason of frequency reduction is attributed to a decrease in stiffness of the nanowire as its temperature increases. Using Eq. (32) based on the FDM solution, we have performed a sensitivity analysis with respect to the piezoelectric voltage, presented in Fig. 1b. Based on this figure, increasing the piezoelectric voltage reduces the frequency of silicon nanowire resonator. Accordingly, the piezoelectric voltage can be used for adjusting the vibration behavior of the nanowire resonator. Figure 2a depicts the effect of dimensionless nonlocal parameter ($(e_0a)/L$) on the frequency behavior of the silicon nanowire resonator. This figure has been plotted by using Eq. (32) based on the FDM solution. As the figure reveals, increasing the nonlocal parameter reduces the frequency of nanowire resonator. From this figure it can be concluded that the effect of nonlocal parameter is critical and it should be taken into account in the frequency analysis of nanoscale resonators such as nanowires.

Figure 2b shows the effect of added mass on the frequency behavior of SiNW. As this figure shows, increasing the mass of added particle reduces the frequency of SiNW resonator. This figure also demonstrates a significant potential of the nanowire

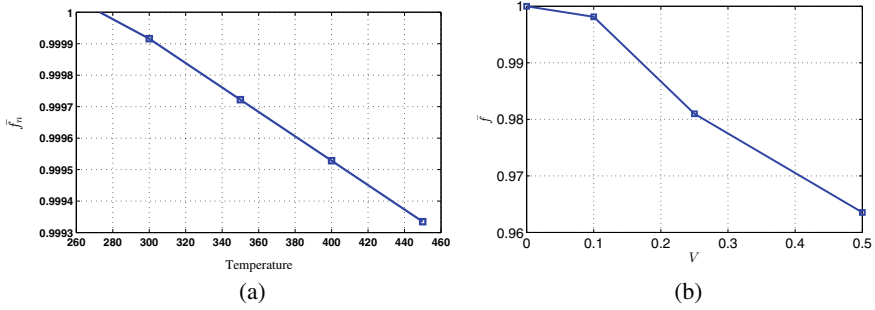


Fig. 1 **a** Effect of temperature on the frequency behavior of SiNW using the FDM **b** effect of piezoelectric voltage on the frequency behavior of SiNW using the FDM

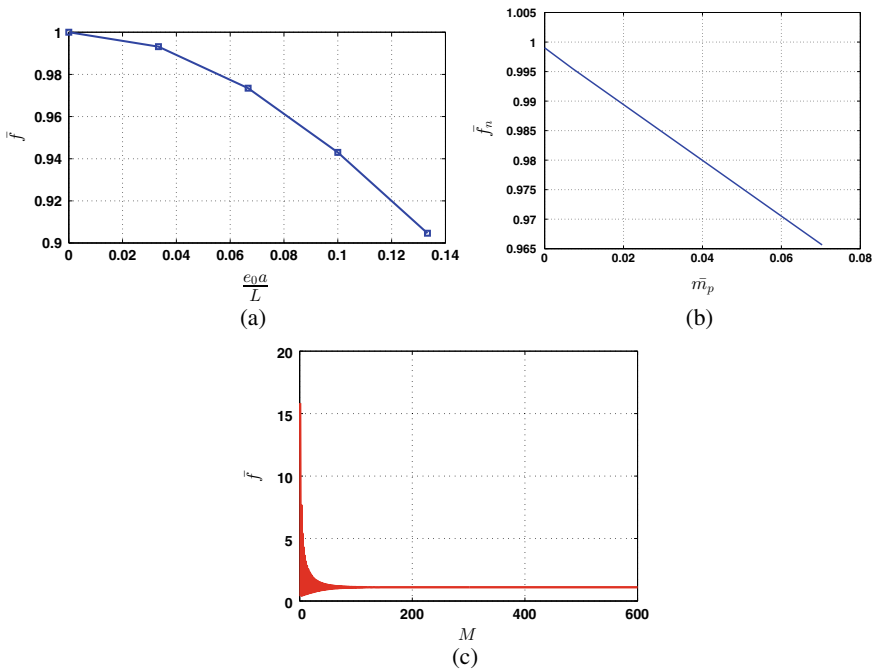


Fig. 2 **a** Effect of dimensionless nonlocal parameter on the frequency behavior of SiNW using the FDM **b** effect of added mass on the frequency behavior of SiNW using the FDM **c** convergence analysis of the frequency response of the FDM

resonator for tiny object detection. In order to investigate the convergence of our numerical solution for the developed model of the silicon nanowire resonator, using the iterative technique in conjunction with the FDM discussed in the context of Eq. (31), we have plotted the obtained nonlinear frequency of each iteration with respect to its corresponding number of iteration, M . Figure 2c shows that by using

the iterative technique for the nonlinear part, we can reach the convergent frequency after just a few iterations with the accuracy of 10^{-4} .

5 Conclusion

In this paper, we presented a numerical solution using the FDM for the vibrations of nanowire resonator with added mass. The mathematical model for the nanowire resonator was developed based on the nonlocal Euler-Bernoulli beam theory, which includes different terms related to thermal variations, electromagnetic fields, surface and nonlocal effects, as well as added mass. It was revealed that the FDM can effectively be used to model nanoresonators and analyze their frequency of oscillations with applications to tiny mass sensing which is critical in biomedicine. It was demonstrated that an increase in temperature, piezoelectric voltage and nonlocal parameter reduces the frequency of oscillations in the nanowire resonator. In addition, adding a tiny particle, in the scale of zeptogram, in the middle of nanowire resonators results in a detectable shift in their frequency.

References

1. Yang, Y.T., Callegari, C., Feng, X.L., Ekinici, K.L., Roukes, M.L.: Zeptogram-scale nanomechanical mass sensing. *Nano Lett.* **6**(4), 583–586 (2006)
2. Joshi, A.Y., Sharma, S.C., Harsha, S.: Zeptogram scale mass sensing using single walled carbon nanotube based biosensors. *Sens. Actuat. A Phys.* **168**(2), 275–280 (2011)
3. Adhikari, S., Chowdhury, R.: Zeptogram sensing from gigahertz vibration: graphene based nanosensor. *Phys. E Low-dimens. Syst. Nanostruct.* **44**(7), 1528–1534 (2012)
4. Patolsky, F., Zheng, G., Lieber, C.: Nanowire sensors for medicine and the life sciences. *Nanomedicine* **1**(1), 51–65 (2006)
5. Norouzzadeh, A., Ansari, R., Rouhi, H.: Nonlinear wave propagation analysis in Timoshenko nano-beams considering nonlocal and strain gradient effects. *Meccanica* **53**(13), 3415–3435 (2018)
6. Barretta, R., Luciano, R., de Sciarra, F.M., Ruta, G.: Stress-driven nonlocal integral model for timoshenko elastic nano-beams. *Eur. J. Mech. A/Solids* **72**, 275–286 (2018)
7. Farajpour, A., Farokhi, H., Ghayesh, M.H., Hussain, S.: Nonlinear mechanics of nanotubes conveying fluid. *Int. J. Eng. Sci.* **133**, 132–143 (2018)
8. Jamshidifar, H., Askari, H., Fidan, B.: Parameter identification and adaptive control of carbon nanotube resonators. *Asian J. Control* **20**(4), 1329–1338 (2018)
9. Fallahpourghadikolaie, R.: Multiscale mathematical modelling of nonlinear nanowire resonators for biological applications. Master's thesis, Wilfrid Laurier University, Waterloo, Canada (2019)
10. Ansari, R., Hosseini, K., Darvizeh, A., Daneshian, B.: A sixth-order compact finite difference method for non-classical vibration analysis of nanobeams including surface stress effects. *Appl. Math. Comput.* **219**(10), 4977–4991 (2013)
11. Krishnan, A., George, G., Malathi, P.: Use of finite difference method in the study of stepped beams. *Int. J. Mech. Eng. Educ.* **26**(1), 11–24 (1998)

12. Mohebbi, A., Dehghan, M.: High-order solution of one-dimensional sine-gordon equation using compact finite difference and dirkn methods. *Math. Comput. Modell.* **51**(5–6), 537–549 (2010)
13. Rao, S.: *Vibration of Continuous Systems*. Wiley (2007)

Contaminant Removal in Ceramic Water Filters by Bacterial Biofilms



Harry J. Gaebler, Jack M. Hughes, and Hermann J. Eberl

Abstract We investigate point-of-use ceramic water filters by reformulating an existing multi-scale biofilm model that has been developed for porous medium applications. The reactor model is described by a stiff system of quasilinear hyperbolic balance laws, which are studied numerically. The model considers processes related to hydrodynamics and transport of a single target contaminant, growth/death of bacteria (both attached biomass inside the filter base, in the form of biofilms, and suspended bacteria), and mass exchange between the biofilm and suspended bacteria via attachment and detachment. With this model, we investigate the influence of water height and refill frequency on the amount and quality of recoverable water.

Keywords Balance laws · Biofilms · Ceramic water filters · Numerical simulation

1 Introduction

In many developing countries access to clean drinking water is not always readily available, resulting in the need to collect water from other locations. By collecting water elsewhere and returning home, there is a high chance that the source water is contaminated [10], often with carbonous organic substrates and/or microbial contaminants. Many point-of-use water filters have been used in the treatment of contaminated source water, with the most widely selected water filter being ceramic water filters (CWF) [13]. CWFs play a vital role in the local treatment of drinking water and are relatively inexpensive to produce. Filters of this type are constructed of clay, sand, and organic material such a rice husks and sawdust [7, 10]. During the

H. J. Gaebler (✉) · J. M. Hughes · H. J. Eberl
University of Guelph, 50 Stone Rd. E., Guelph, Canada
e-mail: gaeblerh@uoguelph.ca

J. M. Hughes
e-mail: jhughe12@uoguelph.ca

H. J. Eberl
e-mail: heberl@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_50

545

firing process, the organic material is burned away and small pore spaces remain in the bed of the filter [10]. These pores help control the flow rate of water (typically 1–3 L/h) through the filter and remove contamination [13].

Over extended periods of use, biofilms begin to form inside these filters. Under certain conditions, biofilm formation can further increase contamination by entering the recoverable water via detachment. One option for effectively inhibiting biofilm growth is to coat the filter in gold or silver nanoparticles [7, 10]. These nanoparticles inhibit the bacterial cells from performing their basic functions, effectively killing off the bacteria in the filter. A laboratory experiment conducted by [7] investigated the effects of silver nanoparticles on biofilm growth by passing contaminated source water through coated and uncoated filters. Results indicated that biofilm growth was inhibited on the filter that was coated with the nanoparticles. The authors of [7] suggest that the growth inhibition was related to the attachment of bacteria to the CWF walls.

In this work, we adapt an existing multi-scale biofilm model described in [5] to investigate how biofilm growth and suspended bacteria contribute to substrate degradation and the quality of recoverable water in a CWF under different hydraulic loading configurations. In this approach, biofilms contribute towards substrate degradation, increasing removal and we investigate their effect on recoverable water.

2 Mathematical Model

In the following model, we consider biofilm growth in a homogeneous porous medium with a well-defined flow direction. A growth limiting substrate and suspended bacteria are transported through the medium by convection and the substrate is consumed by the biomass to promote growth, both on the substratum (in this case the ceramic filter) and in the bulk liquid. Attachment of suspended biomass and detachment of attached biomass are both considered as the biofilm thickness evolves.

The filter is described by parallel non-communicating flow channels of width ε [L], similar to [1, 5]. In the flow direction, each channel is compartmentalized into smaller segments of length ε where mesoscopic processes related to biofilm growth are described following the traditional one-dimensional biofilm model described in [15]. The flow rate in each channel of width ε is the same, but does not remain constant. The flow through the medium is driven by the pressure of the water column in the CWF. The flow process here is in contrast to [5], where the flow rate remained constant and a decrease in the flow path due to biofilm growth resulted in an increase in local flow velocities. A graphical representation of the filter is given in Fig. 1.

The macroscopic reactor is obtained by continuously shrinking the compartment size to zero, i.e. $\varepsilon \rightarrow 0$. The macroscopic reactor is given by the system of stiff quasilinear hyperbolic balance laws

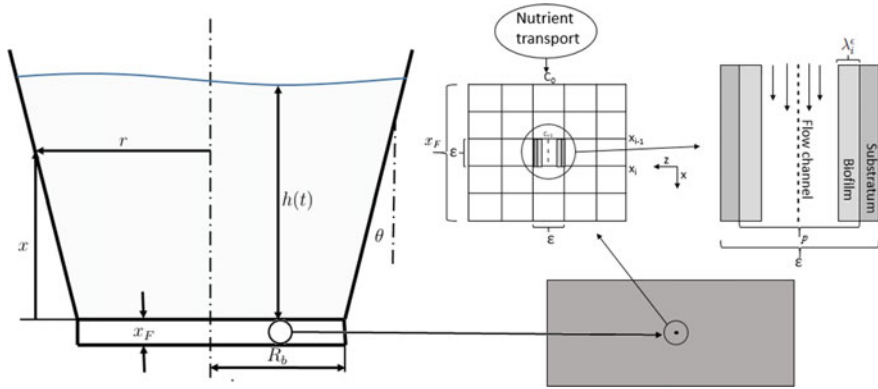


Fig. 1 Schematic of the CWF (left) with the description of the filter in the base of the ceramic pot (bottom-right), the macroscopic compartmentalization (top-middle), and mesoscopic cells of size $\epsilon \times \epsilon$ (top-right)

$$\frac{\partial}{\partial t} \begin{pmatrix} (p - 2\lambda)C \\ (p - 2\lambda)U \\ \lambda \end{pmatrix} + \frac{\partial}{\partial x} \begin{pmatrix} QC \\ QU \\ 0 \end{pmatrix} = \begin{pmatrix} -2J(\lambda, C) - \frac{[p-2\lambda]}{Y_u}g(C)U \\ 2X_\infty d\lambda - 2a[p - 2\lambda]U + g(C)[p - 2\lambda]U \\ \frac{Y}{X_\infty}J(\lambda, C) - k_d\lambda - d\lambda + \frac{a}{X_\infty}[p - 2\lambda]U \end{pmatrix}, \quad (1)$$

where C, U, λ respectively describe substrate concentration [gm^{-2}], suspended bacteria concentration [gm^{-2}], and the unitless biofilm thickness relative to the pore size [-]. In this system, $g(C)$ represents the growth kinetics for suspended bacteria and $J(\lambda, C)$ represents the flux of substrate into the biofilm layer. We model the growth kinetics for suspended bacteria by the commonly used Monod equation [3, 11], which relates the specific growth rate to the concentration of a growth limiting substrate, i.e.

$$g(C) = \frac{\mu_u C}{\kappa_u + C},$$

where μ_u [T^{-1}] is the maximum growth rate for suspended bacteria and κ_u [gm^{-2}] is the suspended bacteria half saturation constant.

Substrate flux into the biofilm layer is determined from the solution to the two-point boundary value problem

$$D \frac{d^2 c}{dz^2} = \frac{X_\infty}{Y_\lambda} \frac{\mu_\lambda c}{\kappa_\lambda + c}, \quad \frac{dc}{dz}(0) = 0, \quad c(\lambda) = C, \quad 0 < z < \lambda, \quad (2)$$

where $c = c(z)$ is the substrate concentration within the biofilm, D [m^2d^{-1}] is the substrate diffusion coefficient inside the biofilm, X_∞ [gm^{-2}] is the biofilm density, Y_λ [-] is the yield coefficient, μ_λ [T^{-1}] is the biofilm maximum growth rate, and κ_λ

$[gm^{-2}]$ is the biofilm half saturation constant. The flux of substrate into the biofilm is given by

$$J(\lambda, C) = D \frac{dc}{dz} \Big|_{\lambda}. \quad (3)$$

For a complete derivation of the macroscopic model, see [5]. All other parameter descriptions are given in Table 1.

Table 1 Parameter values for simulations

Parameter	Symbol	Value	Units	References
Initial water height	h_0	0.2	m	Modified, [12]
Radius of the bottom	R_b	0.1	m	Modified, [12]
Filter thickness	x_F	0.02	m	Modified, [12]
Taper angle	θ	10.0	deg.	Modified, [12]
Hydraulic conductivity	K	0.15	m/d	Calculated
Initial substrate concentration	C_{in}	30.0	g/m^2	[14]
Initial suspended concentration	U_{in}	10.0	g/m^2	[5]
Initial biofilm thickness	λ_{in}	0.0	–	Assumed
Biomass density	X_{∞}	10000.0	gm^{-2}	[14]
Biofilm maximum growth rate	μ_{λ}	6.0	d^{-1}	[14]
Biofilm half saturation constant	κ_{λ}	4.0	gm^{-2}	[14]
Biofilm yield coefficient	Y_{λ}	0.63	–	[14]
Suspended bacteria maximum growth rate	μ_u	6.0	d^{-1}	[14]
Suspended bacteria half saturation constant	κ_u	4.0	gm^{-2}	[14]
Suspended bacteria yield coefficient	Y_u	0.63	–	[14]
Void fraction	p	0.5	–	[5]
Biofilm natural cell death rate	k_d	0.4	d^{-1}	[14]
Detachment coefficient	d	0.5	d^{-1}	[1]
Attachment coefficient	a	0.3	d^{-1}	[5]
Flow rate	Q	Varied	md^{-1}	Calculated
Diffusion coefficient	D	10^{-4}	m^2d^{-1}	[14]

2.1 Pressure Driven Flow

Unlike the model presented in [5], where the flow rate through the medium was constant, we derive pressure driven flow in a CWF with tapered sides (*cf.* Fig. 1). This derivation is similar to the flow derivation in [12], with one fundamental difference, we assume that water leaves the CWF through the bottom of the filter only, i.e. no water is filtered through the sides.

Consider a tapered CWF with a base radius R_b , water height $h(t)$, filter thickness x_F , radius r , which increases with height x , and side taper angle θ . The radius of the filter at height x is given by

$$r = R_b + x \tan \theta. \quad (4)$$

Constructing a mass balance for the volume of water in the filter $V(t)$, we have

$$\frac{dV(t)}{dt} = Q(t), \quad (5)$$

where $Q(t)$ is the volumetric flow rate through the bottom of the CWF given by the Darcy flow equation

$$Q(t) = -\frac{kA(P_2 - P_1)}{\mu x_F}, \quad (6)$$

where k is the permeability of the medium, A is the cross sectional area, $P_2 - P_1$ is the pressure drop, μ is the dynamic viscosity of the fluid, and x_F is the length over which the pressure drop occurs. Assuming the density, ρ , of the liquid remains constant, we have the following relationships,

$$k = \frac{\mu K}{\rho g}, \quad h = \frac{P}{\rho g} \quad (7)$$

where K is the hydraulic conductivity of the filter, which is assumed to be constant, and g is the gravitational constant. Substituting (7) into (6) and expressing the area of the base of the filter as $A = \pi R_b^2$, the flow rate is given by

$$Q(t) = \pi R_b^2 K \frac{h(t)}{x_F}. \quad (8)$$

Based on filter geometry, we also have the relationship

$$\frac{dV(t)}{dt} = -\pi (r_h(t))^2 \frac{dh(t)}{dt}, \quad (9)$$

where $r_h(t) = R_b + h(t) \tan \theta$ is the radius of the filter at height $x = h(t)$. Using (4), (5), (8), and (9) we obtain the expression for the change in water height as

$$\frac{dh(t)}{dt} = -\frac{Kh(t)}{x_F} \left(\frac{R_b}{R_b + h(t) \tan \theta} \right)^2. \quad (10)$$

By first solving (10), the flow rate $Q(t)$ can then be determined via (8).

3 Numerical Treatment

3.1 Numerical Method

To study this system numerically, we use a variable transformation in order to investigate how system (1) progresses over time rather than space. The variable transformation is given by

$$S := (p - 2\lambda)C, \quad W := (p - 2\lambda)U. \quad (11)$$

With the variable transformation (11), the system (1) is written as

$$\frac{\partial}{\partial t} \begin{pmatrix} S \\ W \\ \lambda \end{pmatrix} + \frac{\partial}{\partial x} \begin{pmatrix} \frac{QS}{p-2\lambda} \\ \frac{QW}{p-2\lambda} \\ 0 \end{pmatrix} = \begin{pmatrix} -2J \left(\lambda, \frac{S}{p-2\lambda} \right) - \frac{1}{Y_u} g \left(\frac{S}{p-2\lambda} \right) W \\ 2X_\infty d\lambda - 2aW + g \left(\frac{S}{p-2\lambda} \right) W \\ \frac{Y_\lambda}{X_\infty} J \left(\lambda, \frac{S}{p-2\lambda} \right) - (k_d + d)\lambda + \frac{aW}{X_\infty} \end{pmatrix}. \quad (12)$$

We implement the *Uniformly accurate Central Scheme of order 2* developed in [8] to simulate the system of nonlinear balance laws (12). Implementation of the method is described in [4]. We make one modification to the implementation. In [4], the boundary value problem for the flux is solved using a shooting method with an explicit Runge-Kutta-Fehlberg (RKF) method. As the RKF method in [4] is explicit, it may break down under non-substrate limiting conditions due to the stiffness of the boundary value problem. We adopt a semi-implicit finite difference method and use a fixed point iteration for the nonlinear problem, avoiding potential stiffness issues.

3.2 Implementation

The numerical method was implemented in C and compiled and tested using gcc compilers (gcc version 5.0.0). Simulations were carried out on a standard Linux desktop workstation under Ubuntu 18.04.2. All plots were generated using MATLAB v. 8.6.0.267246 (R2015b).

4 Numerical Simulations

In this section, we simulate the flow through a CWF given in Fig. 1 that has a base radius $R_b = 0.1$ [m], filter thickness $x_F = 0.2$ [m] and taper angle $\theta = 10$ [deg.]. In all simulations the water height is initially $h_0 = 0.2$ [m] and the total volume of water in the filter $V_0 = 0.0087$ [m³] ($V_0 = 8.7$ [L]). The hydraulic conductivity of the filter is calculated from (8), under the assumption the filter is initially capable of filtering 2 [L/h] ($Q_0 = 0.048$ [m³/d]), which is within the common range of filtration rates for CWFs listed in [13].

We begin by considering a new CWF, i.e. one with no established biomass in the bottom of the filter. All model parameters and initial conditions are given in Table 1. The CWF is initially filled to a water height of $h = 0.2$ [m] and refilled every 9 h (at 9 h the water height is approximately 0.02 [m]). This process is completed three times (27 h total) and the results are reported in Fig. 2. Initially, there is no biofilm in the reactor. As time progresses, biofilm begins to form on the substratum inside the filter with the thickest part of the biofilm occurring at the top of the filter, which is consistent with the findings in [5]. Although a biofilm is forming in the filter, there is insufficient biomass to completely filter the substrate at $t = 9$ [h]. The CWF is then refilled back to a height of $h = 0.2$ [m]. At time $t = 3$ [h], $t = 6$ [h], and $t = 9$ [h] after refill, there is again an increase in biofilm thickness and a further decrease in the substrate concentration. This trend continues for the third refill of the filter. This illustrative simulation demonstrates the ability of bacterial biofilms to remove carbonaceous substrate when sufficient bacteria is present. As the CWF is continuously used, more bacteria form inside the filter, which can increase the overall effectiveness of the filter.

Next, we consider a CWF that has an established biofilm in the base of the filter (i.e. $\lambda_0 = 0.001$ [-]). We investigate the effect of filling the CWF in three different ways. Simulations occur over a 9 h period and the CWF can be filled multiple times

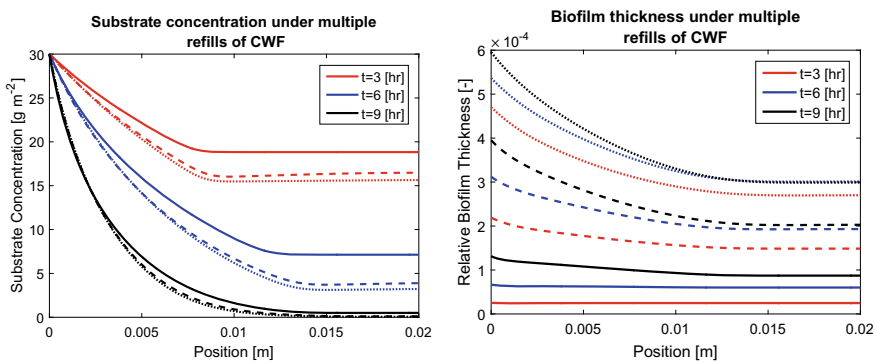


Fig. 2 Substrate concentration (*left*) and relative biofilm thickness (*right*) for three successive refills of the CWF. Solid: initial fill results; dashed: second fill results; dot: third fill results

during that period depending on the availability of untreated water. In the following simulations, we assume that it is feasible that the CWF is refilled up to 3 times. First, the CWF is filled to a height of $h_0 = 0.2$ [m] at $t = 0$ [h] and left for 9 h. Second, the CWF is filled to a height of $h_0 = 0.2$ [m] at $t = 0$ [h] and refilled to a height of $h_0 = 0.2$ [m] at $t = 4.5$ [h]. Third, the CWF is filled to a height of $h_0 = 0.2$ [m] at $t = 0$ [h] and refilled to a height of $h_0 = 0.2$ [m] at $t = 3$ [h] and $t = 6$ [h]. Results are illustrated in Fig. 3.

Results indicate that triple loading yields the most amount of recoverable water at the end of a 9 h filtering period. However, triple loading has the highest concentration of substrate in the outflow after 9 h. This can be attributed to the fact that refilling the CWF increases the amount of water in the filter, which in turn increases the hydraulic pressure forcing more water through the filter. An increased flow rate through the filter decreases the residency time of suspended bacteria, which has a negative impact on substrate degradation (*cf.* Fig. 3a). Under a single substrate loading configuration, the amount of recoverable water is much lower than both the double or triple loading configurations (1.5 and 1.8 times, respectively), but the discharged water has the least amount of substrate (almost zero), suggesting that the triple loading configuration has the smallest removal efficiency.

The largest amount of suspended bacteria occurs under the single loading configuration and the least amount occurs in the triple loading configuration. Since the flow rate is monotonically decreasing over time, suspended bacteria residency time increases, promoting more suspended bacteria growth. However, beyond the first 0.005 m of the filter, substrate concentrations have drastically decreased causing suspended bacteria growth and biofilm thickness to be controlled by attachment, detachment, and lysis processes. Under the conditions of the simulation, the attachment rate is smaller than detachment, which causes a net growth in suspended bacteria and a decrease in biofilm thickness. As very little is known about the biofilm attachment process, which is often chosen out of mathematical convenience [2, 6], we identify the numeric choice of attachment rate as an area requiring further investigation. For the purpose of this work, the attachment value was selected to remain consistent with [5], where the initial reactor biofilm model was developed.

The thickest biofilm forms under the triple loading configuration (*cf.* Fig. 3c). An increased flow rate supplies more nutrients to the biofilm, promoting higher rates of growth (*cf.* Fig. 3c). Additionally, constantly refilling the CWF increases the substrate concentration to the initial contaminant concentration, providing higher concentrations of substrate for bacteria growth. It is important to note that in all three substrate loading configurations the biofilm is the thickest near the top of the filter where substrate availability is the largest and biofilm thickness stratifies near the bottom of the filter. These results are consistent with the findings in [5].

A double loading of the CWF discharges almost as much water as the triple loading (12.2 [L] compared to 14.1 [L], a difference of 14.5%), but has a substantially lower substrate concentration at outflow (5.7 [gm^{-2}] compared to 13.1 gm^{-2} , a difference of 78.9%). This indicates that although the triple loading configuration can filter 1.2 times more water than the double loading configuration, the substrate concentration

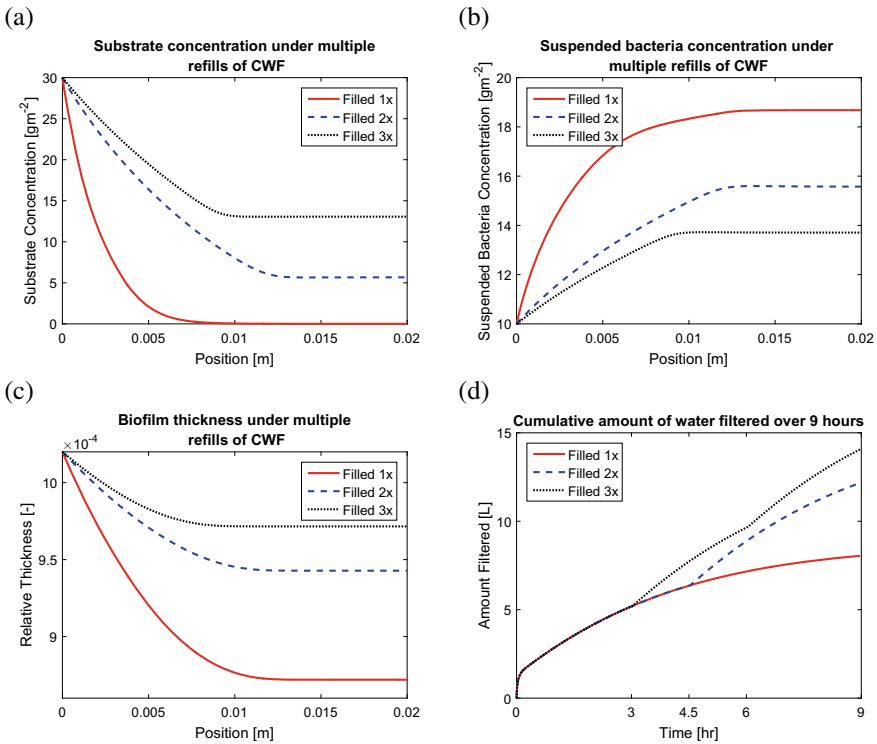


Fig. 3 Solution of (1) with an established biofilm of thickness $\lambda_0 = 0.001[-]$ inside the CWF after 9 h. *a* substrate concentration; *b* suspended bacteria concentration; *c* relative biofilm thickness; *d* amount of recoverable water

in the recoverable water is 2.3 times larger in the triple loading than the double loading. Although the slight decrease in the amount of recoverable water may be advantageous when comparing recoverable water quality.

5 Conclusion

In this work, we adapted a macroscale one-dimensional biofilm growth model developed for porous media applications [5] and applied the resulting system to water filtration in ceramic water filters with pressure driven flow.

Our findings suggest that the amount of recoverable water can be increased by constantly refilling the ceramic filter. Refilling the CWF increases the hydraulic pressure, driving more water through the filter, a result that is consistent with [12]. However, we find that although an increased flow rate leads to more recoverable water, it may have an adverse effect on water quality. Higher flow rates result in

thicker biofilms forming throughout the filter, but also corresponds to less substrate degradation at outflow. While higher flow rates supply the biofilm with more nutrients to promote growth, it also increases the wash out rate of suspended bacteria, which are generally more efficient at removing substrate as they are not subject to concentration gradients [5, 9]. Additionally, we find bacteria in the system do have a presence in the recoverable water. The least amount of bacteria occurs under the single loading configuration and dual/triple loading configurations are comparable.

Overall, this study suggests that increasing the amount of filtered water by strictly increasing the flow rate can have adverse affects on the quality, both in terms of substrate concentrations and bacteria in the recoverable water.

References

1. Abbas, F., Eberl, H.: Investigation of the role of mesoscale detachment rate expressions in a macroscale model of a porous medium biofilm reactor. *Int. J. Biomath. Biostats* **2**(1), 123–143 (2013)
2. Ballyk, M.M., Jones, D.A., Smith, H.L.: Microbial competition in reactors with wall attachment: a mathematical comparison of chemostat and plug flow models. *Microb. Ecol.* **41**, 210–221 (2001)
3. Corapcioglu, M.Y. (ed.): *Advances in Porous Media*, vol. 3. Elsevier, Amsterdam, Netherlands (1993)
4. Gaebler, H.J., Eberl, H.J.: First order versus Monod kinetics in numerical simulation of biofilms in porous media. In: Kilgour, D.M., Kunze, H., Makarov, R., Melnik, M., Wang, X. (eds.) *Recent Advances in Mathematical and Statistical Methods*, pp. 351–361. Springer International Publishing, Switzerland (2018)
5. Gaebler, H.J., Eberl, H.J.: A simple model of biofilm growth in a porous medium that accounts for detachment and attachment of suspended biomass and their contribution to substrate degradation. *Eur. J. Appl. Math.* **29**, 1110–1140 (2018)
6. Jones, D., Kojouharov, H.V., Le, D., Smith, H.: The Freter model: A simple model of biofilm formation. *J. Math. Biol.* **47**, 137–152 (2003)
7. Larimer, C., Islam, M., Ojha, A., Nettleship, I.: Effect of silver nanoparticles on biofilm growth: implications for low-cost ceramic water filters, pp. 1–4. IWA Publishing (2015)
8. Liotta, S.F., Romano, V., Russo, G.: Central schemes for balance laws of reaction type. *SIAM. J. Num. An.* **38**(4), 1337–1356 (2000)
9. Mašić, A., Eberl, H.J.: Persistence in a single species CSTR model with suspended flocs and wall attached biofilms. *Bull. Math. Biol.* **75**(5), 1001–1026 (2012)
10. Nicholson, D.: Carbon fibre reinforcement of ceramic water filters. Master's thesis, University of Guelph, Guelph ON (2012)
11. Rittmann, B.E., McCarty, P.L.: *Environmental Biotechnology: Principles and Applications*. McGraw-Hill, Boston, MA (2001)
12. Schweitzer, R.W., Cunningham, J.A., Mihelcic, J.R.: Hydraulic modeling of clay ceramic water filters for point-of-use water treatment. *Environ. Sci. Technol.* **47**, 429–435 (2012)
13. Sobsey, M., Stauber, C., Casanova, L., Brown, J., Elliott, M.: Point of use household drinking water filtration: a practical, effective solution for providing sustained access to safe drinking water in the developing world. *Environ. Sci. Technol.* **42**, 4261–4267 (2008)
14. Wanner, O., Eberl, H.J., Morgenroth, E., Noguera, D.R., Picioreanu, C., Rittmann, B.E., van Loosdrecht, M.: *Mathematical Modeling of Biofilms*. IWA Publishing, London, England (2006)
15. Wanner, O., Gujer, W.: A multispecies biofilm model. *Biotechnol. Bioeng.* **28**, 314–386 (1986)

Comparison of Fractional-Order and Integer-Order Cancer Tumor Growth Models: An Inverse Approach



Jennifer Lawson and Kimberly M. Levere

Abstract Mathematical modelling of real world phenomena via integer-order differential equation (DE) models has been a popular topic of research for decades. A wide variety of articles have been written in this area and major advancements in model accuracy have been made. Some recent research suggests that fractional-order DEs may more accurately model real world phenomena compared to integer-order counterparts. The development of solution techniques to fractional DEs have been proposed in a number of recent articles. In this paper, we compare fractional-order and integer-order DE models for fitting cancer patient data for tumor growth using fractional DE models. Utilizing actual patient data, we modify three existing integer-order models by instead treating the order of the DE as an unknown parameter. Using a collage-coding inverse problem technique, the order of the DE as well as other parameters in the model are recovered. Finally, results are compared.

Keywords Inverse problem · Fractional differential equation · Cancer tumor growth · Collage theorem · Optimization

1 Introduction

Modelling cancer tumors mathematically is a complex and difficult task. While many ordinary differential equation (ODE) models exist, it is difficult to understand the complex nature of the human body and how it interacts with treatments and the immune system. What is more, the vast majority of these models are of integer-order, perhaps because more common growth and decay models have been derived based on integer-order dynamics.

J. Lawson · K. M. Levere (✉)
University of Guelph, 50 Stone Road E., Guelph, ON N1G 2W1, Canada
e-mail: klevere@uoguelph.ca

J. Lawson
e-mail: jlawso04@uoguelph.ca

© Springer Nature Switzerland AG 2021
D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_51

In this paper, we instead investigate the possibility of utilizing a fractional-order differential equation (FODE) for such modelling (similar explorations of tumor modelling with FODEs have been explored for instance in [10]). While FODEs are perhaps lesser known and studied, great advances in model accuracy have been experienced as discussed in a recent manuscript by Almeida et al. [1], for instance. To this end, in this work, the order of the DE will be left as an unknown parameter to be recovered by an inverse problem technique (collage-coding in our case). In this way, we can investigate if indeed an integer order model is the best fit. In this effort, we will utilize real data from prostate cancer patients who have undergone chemotherapy treatment.

2 Existing Models for Cancer Tumor Growth

A wide variety of models have been produced in the literature regarding different types of cancer and tumors. For instance, in a recent paper by Enderling et al. [7], tumor growth models of many types are discussed from ODE models, to partial differential equation (PDE) models and even discrete models (see [2] for an additional fractional difference model). In this paper we will focus our efforts on ODE models (since the patient data used agrees well with these models). In such models, a first-order ODE measures the change in number of cancer cells over time (or the volume of the tumor), often with a growth term, and perhaps a decay term due to the administration of a drug. There have been many suggestions for an appropriate growth term such as exponential growth, proposed in [4], which observes that the growth of the tumor is proportional to the volume of cells $\frac{dV}{dt} = aV$, where a is the growth rate of the tumor. A logistic model, proposed by Verhulst in [21], observes that there are limits to the size that a tumor can grow based on its surroundings and resources. Here, the growth rate is assumed to decrease in a linear fashion until it is no longer growing when it reaches the carrying capacity, b , $\frac{dV}{dt} = aV \left(1 - \frac{V}{b}\right)$, where a is a growth rate. A so-called linear model initially assumes exponential growth at a rate of $\frac{a}{b}$ that settles in the long term to a constant growth of a as discussed in [6], $\frac{dV}{dt} = \frac{aV}{V+b}$. The name “linear” refers to the fact that growth settles to a constant since indeed this ODE is not linear. Several other growth terms have been explored which we leave the reader to explore, see [8, 14, 17, 20].

Different growth methods seem to better predict cancer tumor growth in different areas of the body; there does not seem to be a one-size-fits-all option. Of course, more intricate biologically motivated models have been proposed as well, see for instance [5]. These authors have a number of manuscripts that seek to model not only the number of cells as a whole, but the number of tumor cells, “natural killer” cells (part of the immune system), and white blood cells. They also include equations for modelling immunotherapy and chemotherapy drug concentrations.

3 Fractional Ordinary Differential Equation Models

One commonality among all of the models listed in the previous section is that they all assume that an integer-order ODE should be used when constructing a model. Since $\frac{dV}{dt}$ defines the change in volume of the tumor over time, certainly this is a reasonable assumption.

More recently, FODEs have been seen to more accurately model certain real-world phenomena. A fractional order model is a refinement of more classical integer-order models that can allow for accuracy not achievable with an integer-order. We know that integer-order derivatives are local in nature and can describe changes in a neighbourhood of a point. On the contrary, fractional-order derivatives are global and can describe changes in an interval. Podlubny provides an explanation in [18] of fractional integrals in terms of measuring the area of so-called “shadows on walls” (parallel to the discussion of areas of rectangles for classical integrals). He also explains an interpretation of fractional derivatives by considering two viewpoints of time. This is motivated, in part, by observations by physicists such as Hawking whom observed the complexity of the concept of time and how it behaves near large bodies as well as how it is interpreted by different observers.

As cancer research is such a complex area of study, it is of interest to investigate if perhaps allowing a fractional order model in place of the aforementioned integer-order models may produce more accurate results.

3.1 A Quick Introduction to Fractional Calculus

Before considering the use of FODE models for improving cancer tumor models, we first need to understand some of the constructs of fractional Calculus. We begin with the definition of a Caputo fractional derivative.

Definition 1 The q th order Caputo fractional derivative of the function $f(x)$ is given by

$${}^C D_a^q f(x) = \frac{1}{\Gamma(q)} \int_a^x (x - t)^{n-q-1} f^{(n)}(t) dt,$$

where $n - 1 < q \leq n$, and $f^{(n)}$ denotes the n th classical derivative of f .

While other definitions for a fractional derivative exist, the Caputo fractional derivative is commonly used in FODEs as it produces physically meaningful initial conditions. Thinking in the opposite direction, we can define the q th order fractional integral.

Definition 2 The q th order fractional integral of the function $f(x)$ is given by

$${}_a J_t^{-q} f(x) = \frac{1}{\Gamma(q)} \int_a^x (x - t)^{q-1} f(t) dt. \tag{1}$$

Using fractional order derivatives we can consider the possibility of FODEs. To develop the necessary theory, we focus on FODEs of the form

$${}^C D_a^q y(x) = f(x, y(x)) \tag{2}$$

$$y^{(k)}(a) = y_0^{(k)}, \tag{3}$$

where $k = 0, \dots, \lceil q \rceil - 1$, f is bounded and Lipschitz continuous in its second argument and (x, y) are in the space $\Omega = [0, \beta] \times [y_0^{(0)} - \alpha, y_0^{(0)} + \alpha]$, for $\alpha, \beta > 0$. Much like its integer-order counterpart, one can prove the existence of a unique solution to (2)–(3) via Banach’s Fixed Point Theorem which we state for completeness.

Theorem 1 (*Banach’s Fixed Point Theorem*) *Let $(X, \|\cdot\|_X)$ be a Banach space and let $T : X \rightarrow X$ be a contractive operator with contraction factor $c \in [0, 1)$. Then there exists a unique fixed point $\bar{x} \in X$ such that $T\bar{x} = \bar{x}$. Moreover, for any $x \in X$, $\|T^{os}x - \bar{x}\|_X \rightarrow 0$ as $s \rightarrow \infty$.*

Proof of this theorem can be found in [22], for instance. When applying Banach’s Fixed Point Theorem to ODEs, a common choice for the contractive, space-preserving operator T is the Picard operator, obtained by integrating the DE (in this case, fractionally) and applying any initial conditions. In terms of the general FODE (2)–(3), the Picard operator is given by

$$(Ty)(x) = \sum_{k=0}^{n-1} \frac{x^k}{k!} y^{(k)}(a) + \frac{1}{\Gamma(q)} \int_0^x (x-t)^{q-1} f(t, y(t)) dt. \tag{4}$$

One can show (see [13]) that provided that f is Lipschitz in its second argument and bounded above in sup norm that indeed the operator in (4) satisfies the hypotheses of Banach’s Fixed Point Theorem and thus, there is a unique solution to (2)–(3).

Much like ODEs, a variety of solution techniques for FODEs exist depending on the form and complexity of the problem. A number of classical ODE techniques exist in FODE theory such as Laplace transforms and power series. Likewise, the exponential that appears in solutions to separable ODEs (for instance) appears as a Mittag Leffler function,

$$E_{\alpha,\beta}(x) = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha k + \beta)},$$

in solutions to similar FODEs. Should the FODE be more complex, a number of numerical techniques have been constructed to find approximate solutions in these settings. For instance, the Adomian Decomposition Method (ADM) can be used to numerically approximate solutions to both linear and nonlinear FODEs. Briefly, this method first applies a fractional integral of order q to the FODE as well as applying any initial conditions. Then by assuming that the solution to the FODE

can be written as the infinite sum of component functions, $\sum_{n=0}^{\infty} u_n(t)$, the so-called Adomian polynomials can be built by substituting this decomposition into the FODE. Using appropriate combinations of subscripts, the Adomian polynomials can be determined and then used to build a recurrence for generating values of the u_n 's. An approximate, truncated solution can then be constructed using a finite number of these component functions, u_n . For a more complete discussion, see [15].

4 Comparison of Integer and Fractional Order Cancer Tumor Models

In Sect. 2, we discussed some of the existing ODE models for cancer tumor growth. We now wish to investigate the possibility of using a fractional-order model to possibly improve results. The rationale is simple: since we are unsure of what order the dynamics should be, we shall leave the order of the DE as an unknown parameter and use an inverse problem technique to identify which order fits experimental data the best. The data that we have used came from a study of patients with castration-resistant prostate cancer, [19]. Since this study involves the administration of a drug to inhibit tumor growth, each of the models discussed in Sect. 2 will have a term of the form $-C_0V$ added, where C_0 represents a decay rate of tumor cells as suggested in [16]. We will utilize a Collage-coding inverse problem technique which we now discuss.

4.1 Inverse Problems via Collage-Coding

The goal of many ODE inverse problems is to find unknown parameter values $\lambda \in \Lambda$ present in the DE such that the solution to the DE involving the chosen parameters, y_λ fits known interpolated data y_{target} well. Mathematically, we wish to minimize the approximation error, subject to $\lambda \in \Lambda$,

$$\min_{\lambda \in \Lambda} \|y_{\text{target}} - y_\lambda\|.$$

In practice, however, approaching this minimization problem head-on can be rather challenging although many methods do exist. Instead, the collage-coding method takes a different approach, attempting to bound this approximation error above by a different distance that is more easily minimized. This upper bound is constructed using the Collage Theorem, a simple consequence of Banach's Fixed Point Theorem which we discussed in Sect. 3.1.

Theorem 2 (*Collage Theorem*) *Let $(X, \|\cdot\|_X)$ be a Banach space and $T : X \rightarrow X$ be a contractive operator with contraction factor $c \in [0, 1)$ and unique fixed point $\bar{y} \in X$. Then*

$$\|y - \bar{y}\|_X \leq \frac{1}{1 - c} \|y - Ty\|_X.$$

The proof of this theorem can be found in [3]. As discussed before, when working with FODEs one choice for the contractive, space-preserving operator T is the Picard operator, (4). Recognizing that $\|y - \bar{y}\|_X$ is just another way to state the approximation error, by minimizing the so-called collage distance $\|y - Ty\|_X$, we can ensure that the approximation error is indeed controlled provided that c is bounded away from 1. This general idea has been applied to a variety of problems, see for instance [11–13].

4.2 Examples

We investigate each of the ODE growth models in Sect. 2, inversely using the data from [19] to approximate the solution to the forward problem, $y(t)$, but leaving the order of the DE, and perhaps other coefficients of the model as unknowns to be found via Collage-coding. Subsequently, using our recovered parameters to construct the model, we will solve the forward problem for y_λ (the recovered solution) and plot this solution together with the given data. All numerical simulations were completed using Maple and some of its built in functions.

Example 1 We consider the possibility of a fractional-order exponential model (now including a decay term) given by

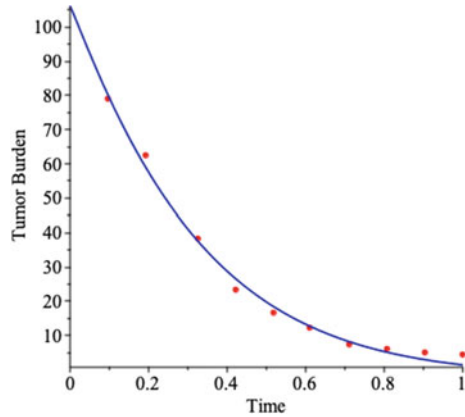
$$\begin{aligned} {}^C D_0^q V &= aV - C_0V \\ V(0) &= V_0 \end{aligned}$$

For the inverse problem, we allow the order of the DE, q , to be unknown (and possibly fractional), as well as the values of a the growth rate, and C_0 the decay rate. The data points were fit with a fourth degree polynomial via least squares and this polynomial was used in the Collage distance in place of the solution function $V(t)$. Using the data, we take the initial volume to be $V_0 = 106.13$. Employing collage-coding using the Picard operator T in (4) we find that the squared collage distance is given by

$$\Delta^2 = \int_0^1 \left(V_{\text{target}}(t) - V_0 - \frac{1}{\Gamma(q)} \int_0^t (t - s)^{q-1} (aV_{\text{target}}(s) - C_0V_{\text{target}}(s)) ds \right)^2 dt$$

Since the parameters appear in a relatively simplistic way in the Collage distance, Maple’s basic “minimize” function can easily solve for each of the parameters in this

Fig. 1 Plot of recovered solution y_λ assuming exponential growth together with given experimental data



example. Certainly though, for more complex problems, more exotic minimization techniques would need to be employed.

As this is truly an inverse problem, the true values of these parameters are unknown and thus we will use the collage distance and approximation error as well as plots to assess the accuracy of our results. The minimal collage error of $\Delta = 0.1361$ is achieved for parameter values of $\lambda = \{q, a, C_0\} = \{1.0544, 0.9990, 4.3585\}$ with an approximation error of 0.1248. In Fig. 1, the recovered solution is plotted together with the experimental data.

We see that indeed the fit is quite strong. The relatively small approximation error indicates that indeed the parameters found indeed produce a solution that agrees closely with the experimental data. In this case, the inverse problem finds that the order of the DE is not far from integer-order at $q = 1.0544$. Since this problem has a closed-form solution involving the Mittag Leffler function, some error was introduced by truncating this infinite sum. This may be alleviated by using a rational approximation of the Mittag-Leffler function that has been shown to have minimal error, see [9] for further details.

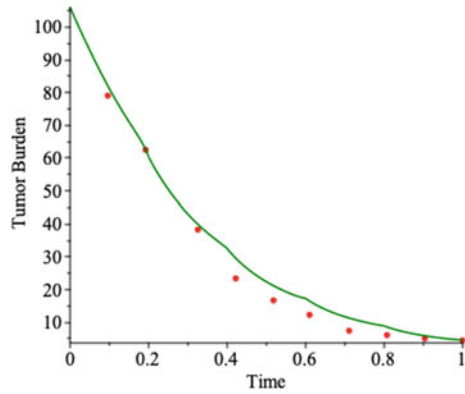
Example 2 We consider the possibility of a fractional-order logistic model (now including a decay term) given by

$$\begin{aligned}
 {}^c D_0^q V &= aV \left(1 - \frac{V}{b}\right) - C_0 V \\
 V(0) &= V_0
 \end{aligned}$$

With the same set up as before, we use collage-coding to recover unknown parameters $\lambda = \{q, a, v, C_0\}$ by minimizing the squared collage distance given by

$$\Delta^2 = \int_0^1 \left(V_{\text{target}}(t) - V_0 - \int_0^t (t-s)^{q-1} \left(a V_{\text{target}}(s) \left(1 - \frac{V_{\text{target}}(s)}{b}\right) - C_0 V_{\text{target}}(s) \right) ds \right)^2 dt.$$

Fig. 2 Plot of recovered solution y_λ assuming logistic growth together with given experimental data



The minimal collage error of $\Delta = 0.0462$ was achieved for parameter values of $\lambda = \{q, a, b, C_0\} = \{0.9689, -6.8997, 299.8911, -2.2048\}$ with an approximation error of 0.8966 Fig. 2.

Once again, the order of the DE recovered is not quite integer. We see that both the growth rate and decay rate are found to be negative, perhaps suggesting that these two terms have reversed roles in this case. Here, the Adomian Decomposition Method was used to find a 16-term approximation to the solution of the DE using the recovered parameters. This truncation together with the fact that this decomposition is only accurate near 0 contribute to some of the error seen in this case. Since the parameters appear in complex ways in the collage-distance, in order to simplify the effort of the minimization scheme (least-squares in this case), Taylor approximations of elements of the integrands were taken introducing additional error in our results Fig. 2.

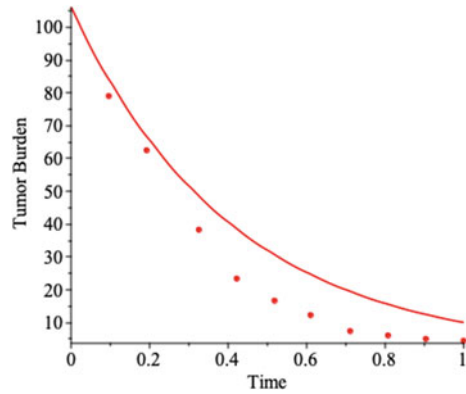
Example 3 Lastly, we consider the possibility of a fractional-order linear model (now including a decay term) given by

$$\begin{aligned}
 {}^c D_0^q V &= \frac{aV}{V + b} - C_0 V \\
 V(0) &= V_0
 \end{aligned}$$

Utilizing the same data as before, with the same initial condition, we utilize the collage-coding method to recover the parameters $\lambda = \{q, a, b, C_0\}$ that minimize the associated collage distance. The minimal collage error of $\Delta = 0.0117$ was achieved for parameter values of $\lambda = \{q, a, b, C_0\} = \{1.0822, 1.7662, -8.8225, 3.5040\}$ and an approximation error of 4.0270.

This model produces the worst approximation error of the examples considered. Much like Example 2, some error was introduced as the result of numerical approximations such as the use of Taylor expansions to simplify minimization, decomposi-

Fig. 3 Plot of recovered solution y_x assuming linear growth together with given experimental data



tion truncation, and the fact that the decomposition method struggles away from 0. The order of the DE recovered is also farthest away from the integer-order.

5 Conclusions and Future Work

While modelling cancer tumor growth is a difficult task, this paper suggests allowing more freedom when it comes to the order of the DE. Since fractional order models allow us to refine existing integer-order models, it may be the case that indeed FODEs can provide us with a granularity not achievable using integer orders. This paper revealed only mild deviations from integer-order, although only very simple models were used and many parameters beyond just the order of the model were also recovered. These models were investigated in part because of the available data. Perhaps the use of less error-inducing FODE solution techniques can reduce some of the errors seen as well. The investigation of more complex cancer tumor growth models may shed additional light on how FODE models (or perhaps even fractional PDE models) might give rise to models that better describe this complex disease. More accurate modeling of chemotherapy involving optimal control should also be investigated.

References

1. Almeida, R., Bastos, N., Monteiro, M.: Modeling some real phenomena by fractional differential equations. *Mathemat. Methods Appl. Sci.* **39**(16), 4846–4855 (2016)
2. Atici, F.M., Sengul, S.: Modeling with fractional difference equations. *J. Mathem. Anal. Appl.* **369**(1), 1–9 (2010)
3. Barnsley, M.F., Ervin, V., Hardin, D., Lancaster, J.: Solution of an inverse problem for fractals and other sets. *Proc. Natl. Acad. Sci.* **83**, 1975–1977 (1985)

4. Collins, V.P., Loeffler, R.K., Tivey, H.: Observations on growth rates of human tumors. *Amer. J. Roentgenol Radium Ther. Nuc. Med.* **78**(5), 988–1000 (1956)
5. de Pillis, L.G., Gu, W., Radunskaya, A.E.: Mixed immunotherapy and chemotherapy of tumors: modeling applications and biological interpretations. *J. Theor. Biol.* **238**(4), 841–862 (2006)
6. Dethlefsen, L.A., Prewitt, J.M.S., Mendelsohn, M.L.: Analysis of tumor growth curves. *J. Nat. Cancer Inst.* **40**(2), 389–405 (1968)
7. Enderling, H., Chaplain, M.A.J.: Mathematical modeling of tumor growth and treatment. *Current Pharmaceutical Design* **20** (2014)
8. Gompertz, B.: On the nature of the function expressive of the law of human mortality, and on a new method of determining the value of life contingencies. *Phil. Trans. Roy. Soc.* **237**, 513–585 (1825)
9. Iyiola, O.S., Asante-Asamani, E.O., Wade, B.A.: A real distinct poles rational approximation of generalized Mittag-Leffler functions and their inverses: applications to fractional calculus. *J. Comput. Appl. Mathem.* **330**, 307–317 (2018)
10. Iyiola, O.S., Zaman, F.D.: A fractional diffusion equation model for cancer tumor. *AIP Advan.* **4**(10), 107–121 (2014)
11. Kunze, H., La Torre, D., Vrscay, E.R.: A generalized collage method based upon the Lax-Milgram functional for solving boundary value inverse problems. *Nonlin. Anal. Theory, Methods Appl.* **71**(12), 1337–1343 (2009)
12. Kunze, H., Vrscay, E.R.: Solving inverse problems for ordinary differential equations using the Picard contraction mapping. *Inverse Probl.* **15**, 745–770 (1999)
13. Levere, K.M., Van De Walker, B.: Solving inverse problems for fractional ODEs via the Collage theorem. *Recent Advances in Mathematical and Statistical Methods*, **259** (2018)
14. Mendelsohn, M.L.: Cell proliferation and tumor growth. Blackwell Scientific Publications, Oxford (1963)
15. Momani, S., Shawagfeh, S.: Decomposition method for solving fractional Riccati differential equations. *Appl. Mathem. Comput.* **182**(2), 1083–1092 (2006)
16. Murphy, H., Jaafari, H., Dobrovolny, H.M.: Differences in predictions of ODE models of tumor growth: a cautionary example. *BMC Cancer* **16**(163) (2016)
17. Patt, H.M., Blackford, M.E.: Quantitative studies of the growth response of the Krebs ascites tumor. *Cancer Res.* **14**(5), 391–396 (1954)
18. Podlubny, I.: Geometric and physical interpretation of fractional integration and fractional differentiation. *Fract. Calculus Appl. Anal.* **5**(4), 367–386 (2002)
19. Scher, I., et al.: Randomized, open-label phase III trial of docetaxel plus high-dose calcitriol versus docetaxel plus prednisone for patients with castration-resistant prostate cancer. *J. Clin. Oncol.* **29**(16), 2191–2198 (2011)
20. Vaidya, V.G., Alexandro, F.J.: Evaluation of some mathematical models for tumor growth. *Int. J. Biomed. Comput.* **13**(1), 19–35 (1982)
21. Verhulst, P.F.: Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Mathématique et physique* **10**, 113–121 (1838)
22. Zeidler, E.: Applied Functional Analysis: Applications to Mathematical Physics. Springer-Verlag, New York (1995)

Numerical Modelling of Drug Delivery in an Isolated Solid Tumor Under the Influence of Vascular Normalization



Mahya Mohammadi, Cyrus Aghanajafi, and Madjid Soltani 

Abstract Mathematical models and numerical methods are used in predicting the various approaches of the cancer treatment process. In the present study, the drug delivery is investigated in three different biological tissues, i.e., tumor, normal, and normalized ones with different transport properties to study the effect of the intensity of vascular normalization on the behavior of interstitial fluid flow and drug distribution. The continuity, momentum (Darcy's law), and convection–diffusion equations in the porous media are solved numerically. Results show that vascular normalization reduces the interstitial fluid pressure and sets up the pressure gradient, which can cause fluid flow throughout the tumor. The drug delivery improvement is shown by solute transport analysis, also. It is concluded that the amount of the maximum rate of drug concentration increases in time by vascular normalization. Moreover, normalization can establish the concentration gradient, which consequently improves the penetration of the drug into the inner parts of the tumor.

Keywords Drug delivery · Vascular normalization · Interstitial fluid pressure · Solute transport analysis

M. Mohammadi (✉) · C. Aghanajafi · M. Soltani (✉)
Department of Mechanical Engineering, Toosi University of Technology, Tehran, K. N, Iran
e-mail: mahya.mohammadi@email.kntu.ac.ir

M. Soltani
e-mail: msoltani@uwaterloo.ca

M. Mohammadi
Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

M. Soltani
Centre for Biotechnology and Bioengineering (CBB), University of Waterloo, Waterloo, ON, Canada

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_52

565

1 Introduction

The focus on cancer treatment has changed in recent years. For the past 10 to 20 years, the main effort was to eliminate cancer cells. By recognizing the vital role of angiogenesis in tumor growth, much effort has been made to improve the antiangiogenic drugs [1]. Disrupting the balance of the angiogenesis and anti-angiogenesis factors is responsible for the migration and proliferation of endothelial cells. The excessive endothelial cells and abnormal cells around the vessels cause the formation of the complicated capillary network with high permeability [2, 3]. The result will be a network that is non-homogeneous and abnormal. This abnormal structure, combined with high density, increases the resistance to flow, which makes blood supply difficult. The high permeability of vessels in lack of effective lymphatic system leads to failure of the pressure gradient between the microvascular and interstitium, and therefore the interstitial pressure goes up dramatically [4–6]. This lack of pressure gradient and consequently high interstitial fluid pressure (IFP) causes the drug delivery disrupted. Antiangiogenic drugs normalize the microvascular network by pruning the immature and inefficient vessels and accordingly improve the drug delivery [7].

Jain et al. [8–10] have conducted extensive and fundamental studies in various cancer-related areas. They considered the interstitium as a porous media and capillary network with simple assumptions. They introduced the IFP as an effective factor, which is studied by the assumption of uniform sources and sinks. Soltani and Chen [11], using the macroscopic view, introduced the critical radius of the tumor and the critical radius of the necrotic region as influential parameters. In another study [12], they investigated the effects of tumor shape and size on drug delivery. Further, on this investigation, Sefidgar et al. [13] studied the effect of shape and size of the tumor on drug delivery for improving the assumption of flowing the drug with fluid. They concluded that drug delivery improves in prolate shape in comparison to other tumor shapes, due to the non-uniformity of IFP. Jain et al. [14] studied the vascular normalization by anti-angiogenesis treatment in an avascular tumor with the governing equations of interstitial flow.

Oztork et al. [15] studied the effect of normalization on the delivery of 100 nm liposomes into an avascular tumor. They found normalization effective at a specific size of the tumor. Steuperaert et al. [16] investigated the intraperitoneal chemotherapy by solving interstitial flow and solute transport equations for various shapes of the tumor. They found that the drug penetrates deeper into the small tumors. Moradi Kashkooli et al. [17] have conducted an image-based mathematical study on the treatment effect of chemotherapy in a solid tumor with heterogeneous microvasculature. In another study by Moradi Kashkooli et al. [18], they investigated the effects of different properties of the drug and some characteristics of tumor microenvironment on the quality of drug delivery by analyzing the treatment efficacy and side effects of chemotherapy.

By the above-mentioned literature review, it is found that the fluid flow analysis has been used in the investigation of the transport phenomenon of biological tissue. However, the drug delivery has spatiotemporal behaviors, so, for finding out the

distribution of the drug over both the time and space, the solute transport equation is added to the fluid flow equation in the current study. In addition, the effect of vascular normalization by the anti-angiogenic factor on drug delivery is examined in more realistic detail by introducing the transport parameters of both fluid flow and solute transport systems to the mathematical model.

2 Materials and Methods

2.1 Governing Equations

2.1.1 Interstitial Fluid Flow

Since the timescale of the transport phenomena is much less than the time of growth of the tumor, physiological parameters can be viewed independently of time. Normal tissue or tumor tissue can be considered as a porous medium. The steady-state incompressible form of continuity equation in the porous media with source and sink of mass is [11];

$$\nabla \cdot \vec{V} = \varnothing_B - \varnothing_L \tag{1}$$

where \vec{V} shows the interstitial fluid velocity (IFV). \varnothing_B indicates the source term and is equal to the flow rate of fluid per unit volume from the blood vessels to the interstitium and vice versa. \varnothing_L is the sink term in the normal tissue and is equal to the flow rate of fluid per unit volume from the tissue to the lymph vessels. The source and sink terms are evaluated through Starling’s law as follows [11];

$$\varnothing_B = \frac{L_p S}{V} (P_B - P_i - \sigma_s (\pi_B - \pi_i)) \tag{2}$$

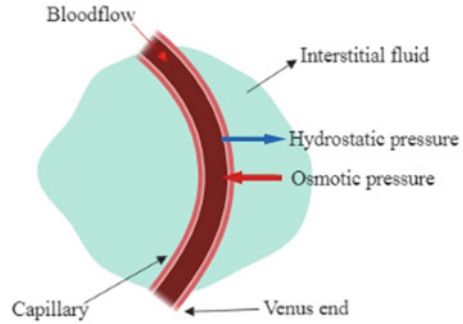
$$\varnothing_L = \frac{L_{pL} S_L}{V} (P_i - P_L) \tag{3}$$

where $\frac{S}{V}$, L_p , P_B , σ_s , π_B , π_i , L_{pL} , and P_L show the surface area per unit volume, hydraulic conductivity of the microvascular wall, vascular pressure, average osmotic reflection coefficient for plasma proteins, osmotic pressure of the plasma, osmotic pressure of the interstitial fluid, hydraulic conductivity of the lymphatic wall, and hydrostatic pressure of the lymphatics, respectively [19].

The simplified momentum equation (Darcy’ law) can be used to describe the fluid flow of the interstitium as follows [13];

$$\vec{V} = -k \nabla P_i \tag{4}$$

Fig. 1 Schematic view of different types of pressure



Where k and P_i are the hydraulic conductivity of the interstitium and the IFP, respectively.

Combining Darcy’s law and the continuity equation when k is constant, results in;

$$-k \nabla^2 P_i = \frac{L_p S}{V} (P_B - P_i - \sigma_s (\pi_B - \pi_i)) - \frac{L_p L S_L}{V} (P_i - P_L) \quad (5)$$

Different types of pressure used in the above equations are shown in Fig. 1.

2.1.2 Solute Transport

The governing equation of the solute transport in biological tissues with a constant diffusion coefficient is as follows [13];

$$\frac{\partial C}{\partial t} = D_{eff} \nabla^2 C - \nabla \cdot (\vec{V} C) + \Phi_b - \Phi_L \quad (6)$$

where C , D_{eff} , Φ_b , and Φ_L represent the concentration, effective diffusion coefficient, the rate of solute transport per unit volume from vessels into the interstitium, and the rate of solute transport per unit volume from the interstitium into lymphatic vessels in normal tissue, respectively. Φ_b , and Φ_L can be considered as follows [13];

$$\Phi_b = \varnothing_B (1 - \sigma_f) C_p + \frac{P S}{V} (C_p - C) \frac{P e}{e^{P e} - 1} \quad (7)$$

$$\Phi_L = \varnothing_L C \quad (8)$$

$\varnothing_B, \varnothing_L, \sigma_f, C_p,$ and P show the fluid flow source, the fluid flow sink, the filtration reflection coefficient, solute concentration in the plasma, and the microvessel permeability coefficient, respectively. Pe is the Peclet number and the related equation is $Pe = \frac{\varnothing_B(1-\sigma_f)V}{PS}$.

According to the Eqs. (6) and (7), the solute transport equation is one-sided coupled to the flow field equation.

2.2 Model Geometry and Boundary Conditions

A homogeneous isolated solid tumor with a specific size was considered in numerical modelling. A schematic view of the computational domain and boundary conditions (BCs) is shown in Fig. 2.

As it is shown in Fig. 2, the no flux boundary condition is applied at the center of the tumor (for $r = 0$) [13]; i.e.,

$$\begin{cases} \nabla P_i = 0 \\ D_{eff} \nabla C + \vec{V} C = 0 \end{cases} \quad (9)$$

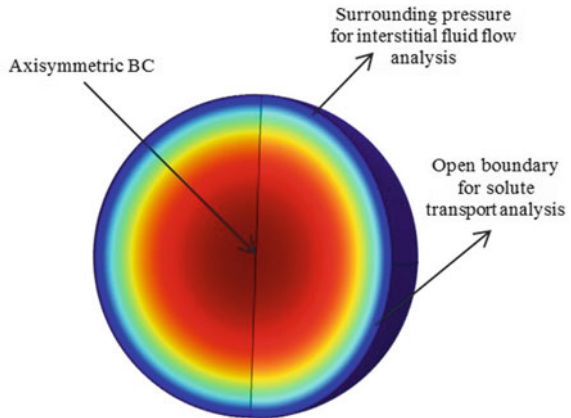
At the edge of the domain (for $r = R$), the IFP is the same as the surrounding pressure, P_{sur} ;

$$P_i = P_{sur} \quad (10)$$

The open boundary condition is used for solute transport analysis at the outer edge of the domain (for $r = R$) [20];

$$-n \cdot \nabla C = 0 \quad (11)$$

Fig. 2 Schematic view of the isolated tumor and boundary conditions



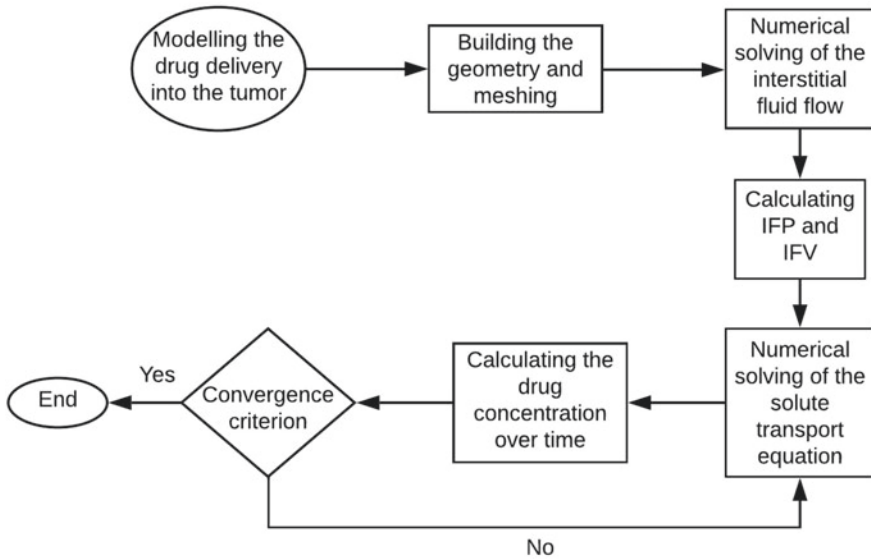


Fig. 3 The simulation flowchart

where n is the normal vector. This boundary condition applies for modeling mass transfer across the boundary where both convective inflow and outflow can occur.

2.3 Solution Procedure

The finite element method (FEM) was used to solve the governing equations, numerically. Quadratic and linear discretization was applied in fluid flow and concentration analyzes, respectively. Newton’s method was set up to solve the equations. The convergence criterion was set to drop the residuals by 6 orders of magnitude. Figure 3 shows the numerical modelling flowchart.

2.4 The Value of Transport Parameters

Parameters of interstitial fluid flow and solute transport properties are listed in Table 1 in different values of α based on the previous studies [8, 14]. D_{eff} was assumed to be 0.75×10^{-8} for normalized tissue. Vascular permeability was considered to decrease in normalized tissue in comparison to the tumor one [15]. α is defined as $\alpha = R\sqrt{\frac{L_p S}{kV}}$ and shows the rate of transport across the vessel wall to the rate through interstitium [14]. R is the tumor radius and is equal to $4mm$ in this study.

Table 1 The value of the parameters of governing equations

Parameter	Description	$1.07 \leq \alpha \leq 2.4$ (normal tissue)	$3.5 \leq \alpha \leq 8$ (normalized tissue)	$7.2 \leq \alpha \leq 17$ (tumor tissue)
$L_p [\frac{cm}{s mmHg}]$	Hydraulic conductivity of the microvascular wall	3.6×10^{-8}	3.7×10^{-7}	1.86×10^{-6}
$k [\frac{cm^2}{s mmHg}]$	Hydraulic conductivity of the interstitium	2.5×10^{-7}	2.5×10^{-7}	2.5×10^{-7}
$\frac{S_f}{V} [\frac{cm^2}{cm^3}]$	Surface area of vessel wall per unit volume of tissue	50 – 250	50 – 250	50 – 250
$P_B [mmHg]$	Vascular pressure	15 – 25	≥ 5.3	5.5 – 34
$\pi_B [mmHg]$	Osmotic pressure of the plasma	20	19.2	19.8
$\pi_i [mmHg]$	Osmotic pressure of the interstitial fluid	10	15.1	17.3
σ_s	Average osmotic reflection coefficient for plasma proteins	0.91	2.1×10^{-3}	8.7×10^{-5}
$D_{eff} [\frac{cm^2}{s}]$ ¹	Effective diffusion coefficient	0.16×10^{-8}	0.75×10^{-8}	2×10^{-8}
σ_f	Filtration reflection coefficient	0.9	0.9	0.9
$P [\frac{cm}{s}]$	Microvessel permeability coefficient	2.2×10^{-8}	10.38×10^{-8}	17.3×10^{-8}

¹The solute transport values were reported for $F(ab')_2$

3 Results and Discussion

The behavior of interstitial flow and solute transport is studied with respect to the different values of α , numerically. In other words, the present work investigates the effect of different transport properties of the vessel wall and the interstitium on IFP, IFV, and drug concentration. Figures 4 and 5 show the IFP and IFV in different values of α . P_{eff} is defined as $P_B - \sigma_s(\pi_B - \pi_i)$ and V_{eff} is the bulk velocity at the margin of the tumor [14].

As it is seen in Fig. 4, in high values of α (tumor tissue), IFP is distributed uniformly in tumor and pressure gradient exists only in a small region in tumor margin. The pressure gradient is established by normalization, which can cause the fluid flow within the interstitium. $\alpha = 5$ is in the normalized region [14]. As it is seen in Fig. 5, the amount of IFV is decreased in tumor margin by normalization. This would result in decreased convection of drugs, and metastatic cancer cells from the tumor margin into the peritumor tissue [14].

Fig. 4 Interstitial fluid pressure in different values of α

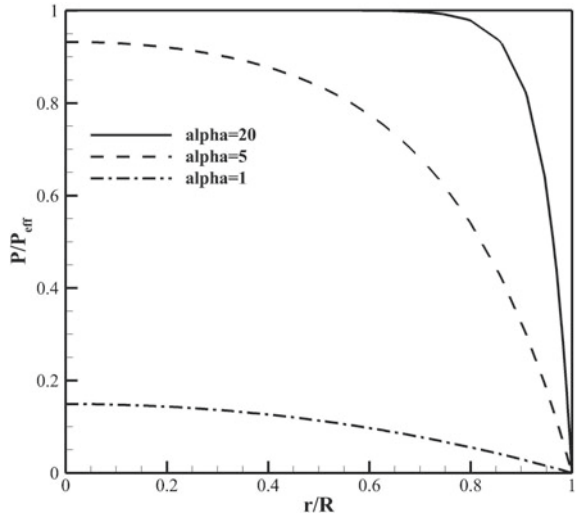
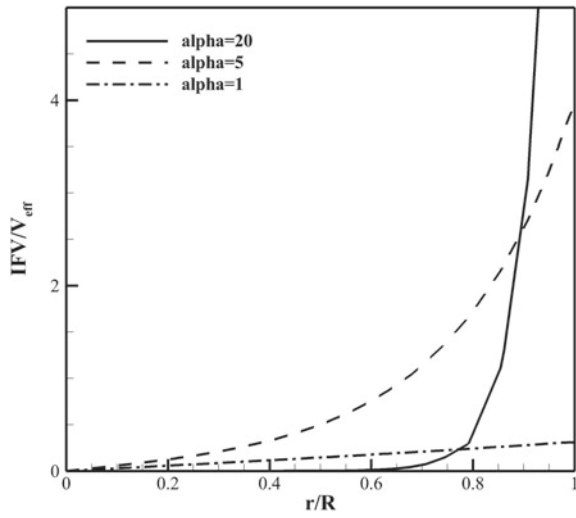


Fig. 5 Interstitial fluid velocity in different values of α



Figures 6–10 show the distribution of drug concentration at different times and different values of α . The continuous injection, which leads to constant plasma concentration, is considered. C_p is considered to be $1\text{ mol}/\text{m}^3$. According to Figs. 6, 7, 8, 9, 10, due to the high permeability of the tumor, the concentration of the drug reaches its maximum value at the initial time after injection, while normalization can change this behavior. The concentration of drug increases in time by vascular normalization.

Fig. 6 Distribution of drug concentration in 1 h post-injection

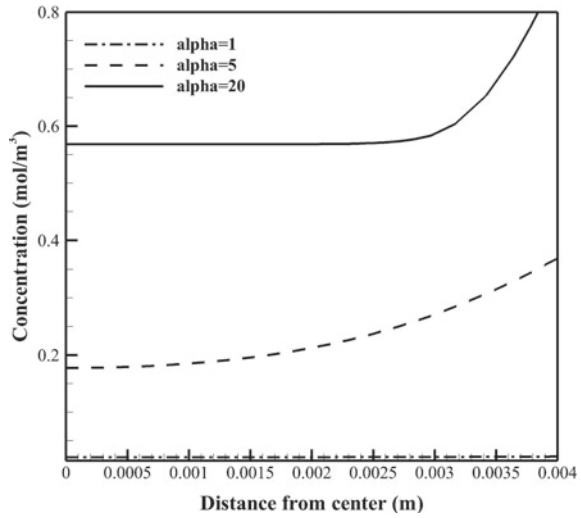
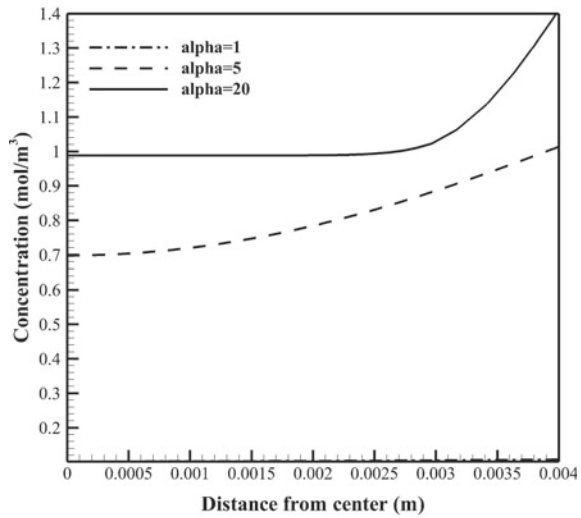


Fig. 7 Distribution of drug concentration in 5 h post-injection



In a high value of α ($\alpha = 20$), the drug has a uniform distribution within the tumor, and only in the region around the tumor, there is a concentration gradient. The concentration gradient is established by normalization that facilitates drug penetration into the inner parts of the tumor. The concentration gradient exists in low values of α throughout the interstitium, but due to the difference of scale, it is hard to show in the figures.

Figure 11 shows the average non-dimensionalized concentration for different values of α over time. It is obvious that the concentration reaches its maximum value

Fig. 8 Distribution of drug concentration in 10 h post-injection

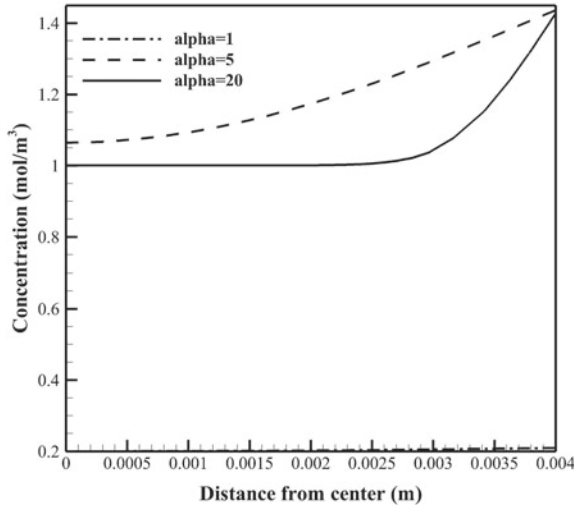
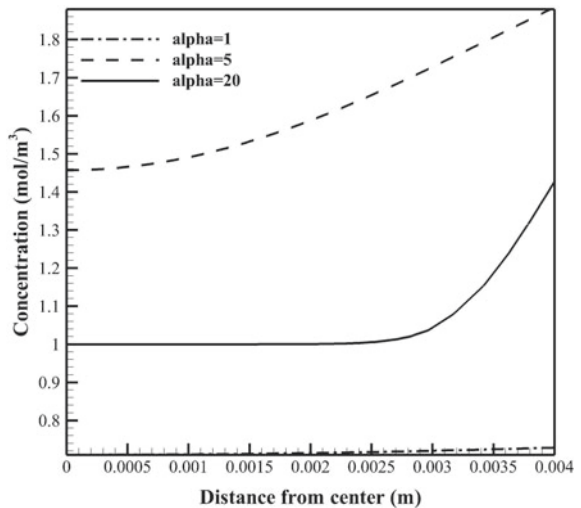


Fig. 9 Distribution of drug concentration in 40 h post-injection



in tumor tissue at the initial time. Normalizing the tumor corrects this behavior. Moreover, the maximum rate of drug concentration increases by normalization.

The drug concentration increases in normal tissue by continuing the injection. The maximum concentration in normal tissue is highest.

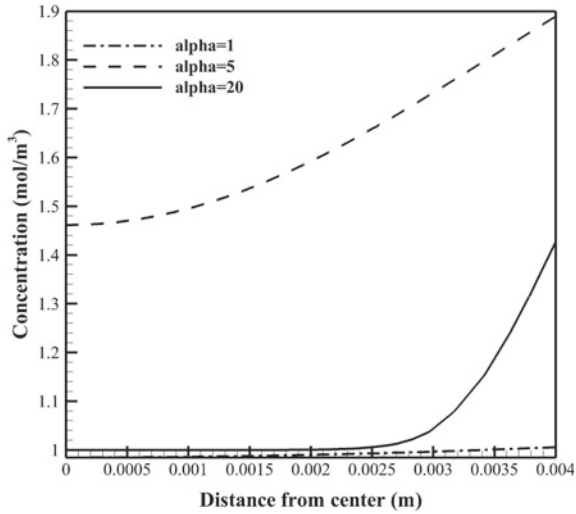


Fig. 10 Distribution of drug concentration in 60 h post-injection

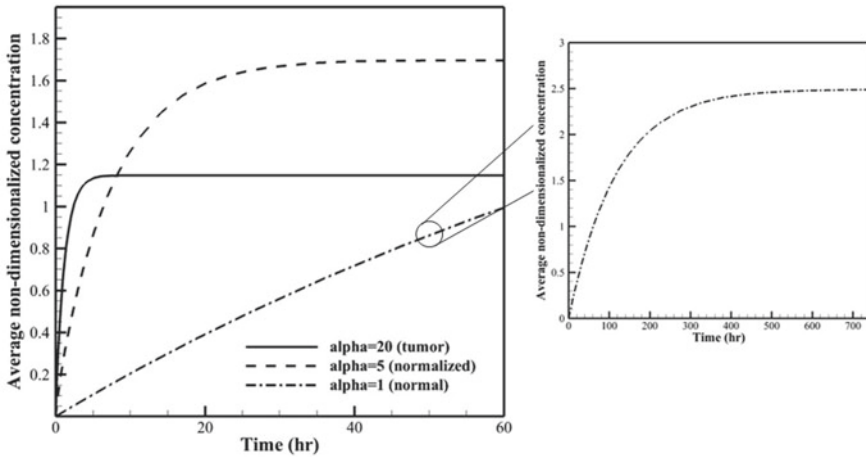


Fig. 11 The average non-dimensionalized concentration over time

4 Conclusion

In this research, the effect of a few improvements in the transport properties of the vessel wall and interstitium on drug delivery is addressed by simulating the interstitial fluid flow and drug concentration. It is found that the interstitial fluid pressure decreases by normalizing the tumor microvessel network. Moreover, normalization

reduces the interstitial fluid velocity at the tumor boundary, which can lower the probability of the occurrence of the metastases. It is realized that the drug can penetrate the tumor by establishing the concentration gradient induced by vascular normalization. The increase of drug concentration through time after normalization is another improvement in delivering the therapeutic agents into the solid tumor.

References

1. Tozer, G.M.: Measuring tumour vascular response to antivascular and antiangiogenic drugs. *British J. Radiol.* **76**(1), 23–35 (2003)
2. Baish, J.W., Jain, R.K.: Fractals and cancer. *Can. Res.* **60**(14), 3683–3688 (2000)
3. Dreher, M., Michelich, L.W., Dewhirst, C., Yuan, F., Chilkoti, A.: Tumor vascular permeability, accumulation, and penetration of macromolecular drug carries. *J. Natl Cancer Inst.* **98**(5), 335–344 (2006)
4. Sefidgar, M., Soltani, M., Raahemifar, K., Sadeghi, M., Bazmara, H., Bazargan, M., Mousavi Naeenian, M.: Numerical modeling of drug delivery in a dynamic solid tumor microvasculature. *Microvasc. Res.* **99**, 43–56 (2015)
5. Soltani, M., Chen, P.: Numerical modeling of interstitial fluid flow coupled with blood flow through a remodeled solid tumor microvascular network. *PLoS ONE* **8**(6), e67025 (2013)
6. Jain, R.K.: The next frontier of molecular medicine: delivery of therapeutics. *Nat. Med.* **4**(6), 655–657 (1998)
7. Jain, R.K.: Normalizing tumor vasculature with anti-angiogenic therapy: A new paradigm for combination therapy. *Nat. Med.* **7**(9), 987–989 (2001)
8. Baxter, L.T., Jain, R.K.: Transport of fluid and macromolecules in tumors. (I) Role of interstitial pressure and convection. *Microvasc. Res.* **37**(1), 77–104 (1989)
9. Baxter, L.T., Jain, R.K.: Transport of fluid and macromolecules in tumors. (II) Role of heterogeneous perfusion and lymphatics. *Microvasc. Res.* **40**(2), 246–263 (1990)
10. Baxter, L.T., Jain, R.K.: Transport of fluid and macromolecules in tumors. (III) Role of Binding and Metabolism. *Microvasc. Res.* **41**(1), 5–23 (1991)
11. Soltani, M., Chen, P.: Numerical modeling of fluid flow in solid tumors. *PLoS ONE* **6**(6), e20344 (2011)
12. Soltani, M., Chen, P.: Effect of tumor shape and size on drug delivery to solid tumors. *J. Biomed. Eng.* **6**(4), (2012)
13. Sefidgar, M., Soltani, M., Raahemifar, K., Bazmara, H., Mousavi Naeenian, M., Bazargan, M.: Effect of tumor shape, size, and tissue transport properties on drug delivery to solid tumors. *J. Biomed. Eng.* **8**(12), (2014)
14. Jain, R.K., Tong, R., Munn, L.L.: Effect of vascular normalization by antiangiogenic therapy on interstitial hypertension, peritumor edema, and lymphatic metastasis: insights from a mathematical model. *Can. Res.* **67**(6), 2729–2735 (2007)
15. Oztork, D., Yonucu, S., Yilmaz, D., Unlu, M.B.: Influence of vascular normalization on interstitial flow and delivery of liposomes in tumors. *Phys. Med. Biol.* **60**(4), 1477–1496 (2015)
16. Steuperaert, M., Labate, G.F.D., Debbaut, C., Wever, O.D., Vanhove, C., Ceelen, W., Segers, P.: Mathematical modeling of intraperitoneal drug delivery: simulation of drug distribution in a single tumor nodule. *Drug Delivery* **24**(1), 491–501 (2017)
17. Moradi Kashkooli, F., Soltani, M., Rezaeian, M., Taatizadeh, E., Hamed, M.H.: Image-based spatio-temporal model of drug delivery in a heterogeneous vasculature of a solid tumor-computational approach. *Microvasc. Res.* **123**, 111–124 (2019)
18. Moradi Kashkooli, F., Soltani, M., Hamed, M.H.: Drug delivery to solid tumors with heterogeneous microvascular networks: novel insights from image-based numerical modeling. *Eur. J. Pharm. Sci.* **151**, 105399 (2020)

19. Soltani, M.: Numerical modeling of drug delivery to solid tumor microvasculature. PhD Dissertation, University of Waterloo (2013)
20. Kim, H.J., Kim, W.: Method of tumor volume evaluation using magnetic resonance imaging for outcome prediction in cervical cancer treated with concurrent chemotherapy and radiotherapy. *Radiat. Oncol. J.* **30**(2), 70–77 (2012)

Quantitative Study of the Coupling Among Cardiovascular System, Lymphatic System and Interstitial Space



Nicholas Mattia Marazzi, Virginia H. Huxley, Riccardo Sacco, and Giovanna Guidoboni

Abstract Tissue interstitial pressure plays a crucial role in maintaining fluid balance in the body. Despite solid experimental evidence of subatmospheric interstitial pressures, the mechanisms leading to the observed negative pressure values have yet to be elucidated fully. The present work addresses this issue theoretically by coupling, for the first time, the cardiovascular and lymphatic circulations within a single closed-loop model. Two model versions are compared, in which lymph formation results solely from a difference in hydrostatic pressure (Model 1), or is the result of the combined action of hydrostatic pressure differences and other mechanisms, such as oncotic pressure gradients and muscle motion (Model 2). Simulations indicate that hydrostatic mechanisms fail to yield negative interstitial pressures, and that suction effects due to lymphatic pumping promote lymph formation without yielding negative interstitial pressures.

Keywords Mathematical modeling · Cardiovascular system · Lymphatic system · Interstitial space

N. M. Marazzi (✉)

Department of Electrical Engineering and Computer Science,
University of Missouri 201 Naka Hall, Columbia, MO 65211, USA
e-mail: marazzin@mail.missouri.edu

V. H. Huxley

Department of Medical Pharmacology and Physiology, University of Missouri,
One Hospital Drive, MA415 Medical Sciences Building, Columbia, MO 65212, USA
e-mail: huxleyv@health.missouri.edu

R. Sacco

Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32,
20133 Milano, Italy
e-mail: riccardo.sacco@polimi.it

G. Guidoboni

Department of Electrical Engineering and Computer Science, Department of Mathematics,
University of Missouri, 201 Naka Hall, Columbia, MO 65211, USA
e-mail: guidobonig@missouri.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_53

1 Introduction

Interstitial spaces within tissues play a crucial role in maintaining fluid balance in the human body [14]. Several experimental measurements have led to the widely accepted concept of subatmospheric pressure in the interstitial space and its crucial role in fluid homeostasis [16]. Alterations in the interstitial pressure are associated with many pathological conditions, including heart disease and lymphedema [34].

A force withdrawing fluid from the interstitium has been hypothesized as a major cause for the negative values of interstitial pressures [15]. Recently, Jamalian et al. provided a first demonstration of a suction effect by combining ex-vivo experiments with mathematical modeling [16]. However, these experiments were performed on the collecting lymphatics and did not focus on whether and to what extent the suction effect contributed to the establishment of subatmospheric pressure levels in the interstitial space. The present work constitutes a first step towards addressing this issue from the theoretical viewpoint by means of a computational model.

To study how interstitial pressure is established, it is necessary to connect the cardiovascular system, the lymphatic system and the interstitial space within the same model. Several mathematical models have been proposed for the cardiovascular function, see [32] for a review. To date, however, mathematical modeling of the lymphatic system remains at its early stages. Most of the available models focus on the contracting element of the lymphatic system, called lymphangion [12, 16, 25, 27], while few works have targeted the initial lymphatics [10, 11, 26]. Interstitial-fluid volume regulation has been studied by means of a closed loop feedback system [9]. Continuum models have been proposed to study the coupling between capillary filtration and interstitial fluid [6] and interstitial fluid and initial lymphatics [28]. Most notably, all the aforementioned studies share the common feature of focusing on single components involved in the maintenance of the fluid balance. To the best of our knowledge, a modeling effort aimed at connecting the multiple components has yet to be attempted. A first step in this direction is presented in this paper.

2 Methods

The electric analogy to fluid flow is used to model the circulation in the cardiovascular and lymphatic systems. In this framework, electric charges represent blood volumes, electric currents represent volumetric flow rates and electric potentials represent fluid pressures. In the present work, two versions of closed-loop models are proposed and compared, each version consisting of a network of electrical elements arranged in three interconnected compartments representing the heart, the systemic circulation and the extravascular circulation, as depicted in Fig. 1. The two models, denoted by *Model 1* and *Model 2*, while are based on the same description for the cardiovascular system and lymphatic return, differ with respect to the connections between interstitial space and lymphatic system.

Cardiovascular system. This block comprises heart and systemic and pulmonary circulations. The pumping action of the heart is modeled via a voltage source and a variable capacitor connected in series reproducing the chemical activation and the time-varying elastance of the ventricles [13, 17]. The heart valves are modeled as ideal switches [13, 17]. The systemic circulation includes arteries, arterioles, capillaries, venules and veins. A combination of resistors, inductors and capacitors (RLC) describes the flow through arteries and arterioles [18, 22], whereas a linear resistor is introduced to model viscous pressure losses due to fluid movement in the capillary beds [18, 19]. Both venules and systemic veins are modeled with a resistor and a capacitor accounting for their elastic behavior [19, 22]. The pulmonary circulation is described as a RLC combination as proposed by Avanzolini et al. [17].

Filtration and interstitial space. The fluid filtration process is modeled through a parallel configuration of a resistor and a constant current source, denoted by R_{11} and I_1 in Fig. 1. The resistor describes the contribution due to hydrostatic pressure, while the current source represents the oncotic pressure inward flux due to gradients in protein concentration. The compliant behavior of the interstitial space is modeled via the linear capacitor C_{12} . Within physiological ranges of volumes and pressures, the interstitial compliance is expected to behave linearly, as shown by the volume-pressure relationship reported in Aukland and Reed [20] and Wigg and Swartz [21].

Lymphatic pumping and return. The lumped modeling of lymphangions is based on the analogy of behavior with the heart [8, 23, 24], where the contractile function is described via an electrochemical activation function and a time-varying elastance [22]. Specifically, the voltage source U_{LY} accounts for spontaneous contractility and the capacitor E_{LY} describes compliant behaviors. The unidirectional flow is guaranteed by the presence of ideal switches that represent secondary valves.

Connection between interstitial space and lymphatic system. The description of this block differs between models 1 and 2 in terms of the assumptions regarding the mechanisms that govern the passage of fluid from the interstitial space to the lymphatic system (lymph formation). Specifically:

- *Model 1:* lymph formation is assumed to be solely due to a difference in hydrostatic pressure. As a consequence, this block is modeled via a linear resistor, whose resistance is denoted by R_{13} in Fig. 1;
- *Model 2:* lymph formation is assumed to result from the combined action of (i) hydrostatic pressure gradients, which are modeled by a linear resistor R_{13} ; and (ii) other mechanisms that are independent of hydrostatic pressure differences, such as oncotic pressure differences and muscle motion, which are modeled by a current source denoted by I_2 .

Model parameter values are summarized in Table 1. Whenever possible, we have selected values within intervals reported in the physiological and computational literature. However, due to both the novelty of the present work and the uncertainty in the mechanisms governing the interstitial and lymphatic interaction, some model parameters have required special attention, as discussed below.

Pressure in the lymphatic vessels. A reference value for the lymphatic pressure is needed to calibrate the lumped parameters pertaining to the lymphatic system.

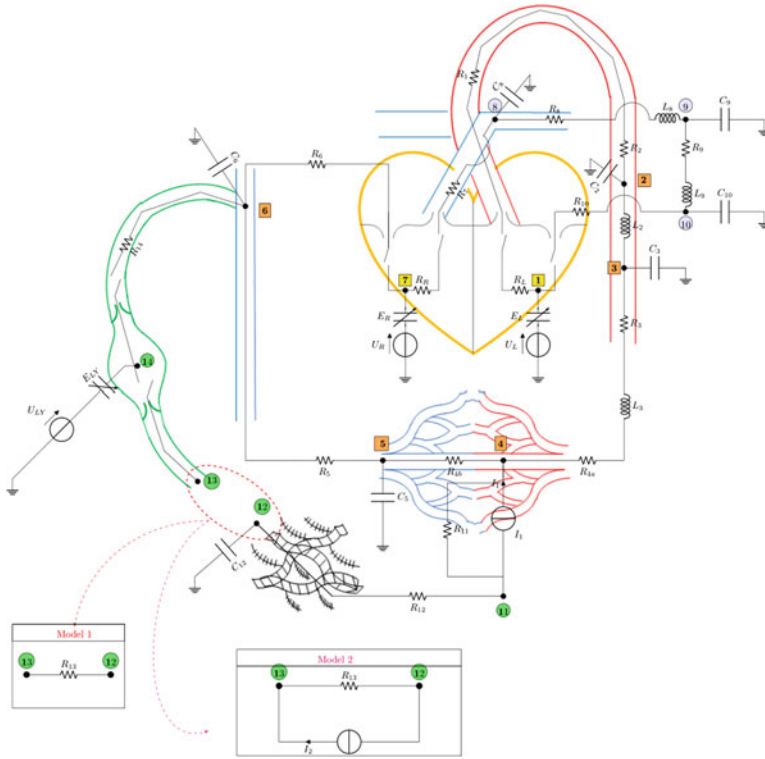


Fig. 1 Closed-loop model including the cardiovascular system, the lymphatic system and the interstitial space. The formulation comprises four different anatomical compartments, whose nodes have been marked with different notations: (i) heart (yellow squares); (ii) the systemic circulation (orange squares); (iii) the pulmonary circulation (blue circles); (iii) the lymphatic circulation (green circles). Model 1 and Model 2 represent two different configurations for the connection between interstitial space and lymphatic system that have been considered and compared in this work

To this end, we considered the measurements in the subcutaneous lymph vessel conducted by Olszweski and Engeset [5], where a mean value of 14.35 ± 9.5 mmHg was reported, and the study performed in human lymphatic leg by Unno et al. [4], where a mean peak pressure value of 25 ± 16.7 mmHg was reported.

Interstitial conductance and compliance. Experimental data for the characterization of interstitial conductance and compliance are currently lacking. Hence, in the present work we have explored how different choices in the effective hydraulic resistance R_{12} and the effective compliance C_{12} would affect the interstitial pressure.

Lymphatic resistance. The measurements of the lymphatic resistance reported in the literature vary in a range of several orders of magnitude. For instance, Papp et al. provided an estimate of the lymphatic resistance measuring flow and pressure in different anatomical locations in dogs [2]. The lymphatic resistance was computed as $10.67 \text{ mmHg s cm}^{-3}$ in the lymph trunk, whereas a value of $3240 \text{ mmHg s cm}^{-3}$ was

Table 1 Summary of model parameters. The parameters indicated with * have been calibrated in this work

<i>Heart</i>		<i>Cardiovascular circulation</i>	
Parameter	Ref	Parameter	Ref
ULO = 50 mmHg	[22]	$R_1 = 0.003751 \text{ mmHg s cm}^{-3}$	[22]
ELD = 0.1 mmHg cm ⁻³	[22]	$R_2 = 0.0675 \text{ mmHg s cm}^{-3}$	[22]
ELS = 1.375 mmHg cm ⁻³ s ⁻¹	[22]	$R_3 = 0.75 \text{ mmHg s cm}^{-3}$	[22]
$R_L = 0.08 \text{ mmHg s cm}^{-3}$	[22]	$R_{4a} = 0.155 \text{ mmHg s cm}^{-3}$	[22]
$T_{c,L} = 0.8 \text{ s}$	[22]	$R_{4b} = 0.155 \text{ mmHg s cm}^{-3}$	[22]
$T_{s,L} = 0.4 \text{ s}$	[22]	$R_5 = 0.125 \text{ mmHg s cm}^{-3}$	[22]
URO = 24 mmHg	[22]	$R_6 = 0.003751 \text{ mmHg s cm}^{-3}$	[22]
ERD = 0.03 mmHg cm ⁻³ s ⁻¹	[22]	$C_2 = 0.21968142 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]
ERS = 0.328 mmHg cm ⁻³ s ⁻¹	[22]	$C_3 = 1.46 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]
$R_R = 0.0175 \text{ mmHg s cm}^{-3}$	[22]	$C_5 = 3.2 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]
$T_{c,R} = 0.8 \text{ s}$	[22]	$C_6 = 8 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]
$T_{s,R} = 0.4 \text{ s}$	[22]	$L_2 = 0.000825 \text{ mmHg s}^2 \text{ cm}^{-3}$	[22]
$L_3 = 0.0036 \text{ mmHg s}^2 \text{ cm}^{-3}$	[22]		
<i>Pulmonary circulation</i>		<i>Lymphatic system and interstitial space</i>	
Parameter	Ref	Parameter	Ref
$R_7 = 0.003751 \text{ mmHg s cm}^{-3}$	[22]	$R_{11} = 3.03 \text{ mmHg s cm}^{-3}$	[33]
$R_8 = 0.03376 \text{ mmHg s cm}^{-3}$	[22]	$R_{12} = 500 \text{ mmHg s cm}^{-3}$	*
$R_9 = 0.1013 \text{ mmHg s cm}^{-3}$	[22]	$R_{13} = 12.5 \text{ mmHg s cm}^{-3}$	*
$R_{10} = 0.003751 \text{ mmHg s cm}^{-3}$	[22]	$R_{14} = 180 \text{ mmHg s cm}^{-3}$	*
$C_9 = 2.67 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]	$C_{12} = 10 \text{ cm}^3 \text{ mmHg}^{-1}$	*
$C_{10} = 46.7 \text{ cm}^3 \text{ mmHg}^{-1}$	[22]	$I_1 = 6.5 \text{ cm}^3 \text{ mmHg}^{-1}$	[33]
$L_8 = 0.00075 \text{ mmHg s}^2 \text{ cm}^{-3}$	[22]	$I_2 = 0.3 \text{ cm}^3 \text{ s}^{-1}$	*
$L_9 = 0.00308 \text{ mmHg s}^2 \text{ cm}^{-3}$	[22]	ELYD = 0.0084 mmHg cm ⁻³	*
		ELYS = 0.42 mmHg cm ⁻³	*
		ULYO = 2 mmHg	*
		$T_{c,L,Y} = 10 \text{ s}$	[1]
		$T_{s,L,Y} = 0.5 \text{ s}$	*

calculated in the submandibular lymph node. The resistance of the extrapulmonary lymph vessel was computed as 1839 mmHg s cm⁻³ by Drake et al. [3]. We have adopted a value of 12 mmHg s cm⁻³ for R_{13} , since the initial lymphatics are expected to provide the smaller resistance to lymph flow [31]. For the resistance R_{14} of the collecting lymphatics, a value of 180 mmHg s cm⁻³ has been adopted.

3 Results

Models 1 and 2 have been implemented and solved in OpenModelica [30]. Then, the simulation results were post-processed using Matlab [29]. The results reported in this section correspond to the last of 16 simulated lymphatic contraction cycles.

Pressures in the cardiovascular system. Figure 2a–c reports the pressure waveforms in arteries, arterioles and veins, corresponding to nodes 2, 3 and 6 in Fig. 1. The pressure values computed via models 1 and 2 are compared with the expected physiological band reported by Guyton [33]. Figure 2d, e reports the pressure waveform in capillaries and venules, corresponding to nodes 4 and 5 in Fig. 1. In both compartments, which are central for our investigation, the simulated values are compared with (i) the theoretical results obtained by Mueller and Del Toro [19] via a multiscale cardiovascular model and (ii) the physiological ranges reported by Guyton [33]. Overall, the predictions of models 1 and 2 regarding blood pressures are consistent with physiological expectations in a healthy human.

Lymphatic function and interstitial pressure. Figure 3 reports the time profile of interstitial pressure P_{12} , which corresponds to node 12 in Fig. 1, as the effective interstitial hydraulic resistance R_{12} is varied in Model 1. The results show that higher values of R_{12} result in lower mean values of P_{12} , see Fig. 3b and Table 2, and smaller

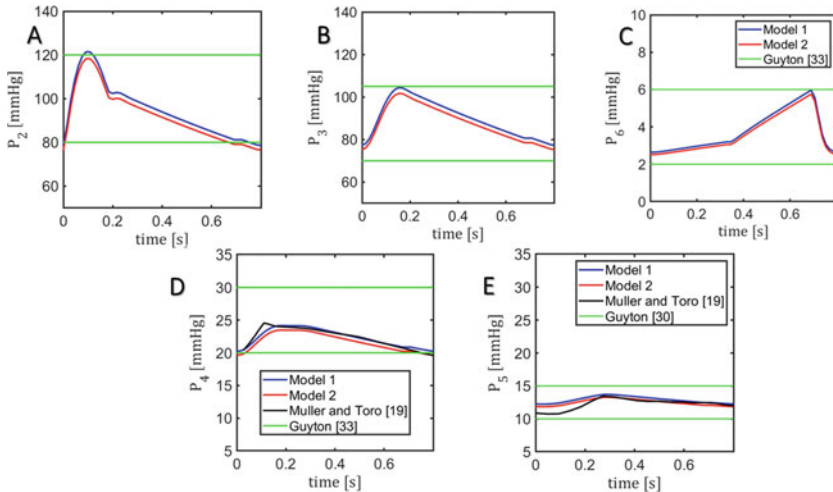


Fig. 2 Pressure waveforms simulated in (A) arteries (P_2), (B) arterioles (P_3) and (C) veins (P_6) are plotted over the period of one cardiac cycle (0.8s). The pressures simulated with Model 1 (blue curve) and Model 2 (red curve) are shown to be within the same order of magnitude as those reported in the by Guyton [33] (green band). Pressure waveforms simulated in (D) capillaries (P_4) and (E) venules (P_5) obtained via Model 1 (blue curve) and Model 2 (red curve) are compared with the theoretical results obtained by Mueller and Del Toro [19] (red curve). The simulated pressure are in good agreement with those obtained by Mueller and Del Toro and fall within the physiological range (green band) reported by Guyton [33]

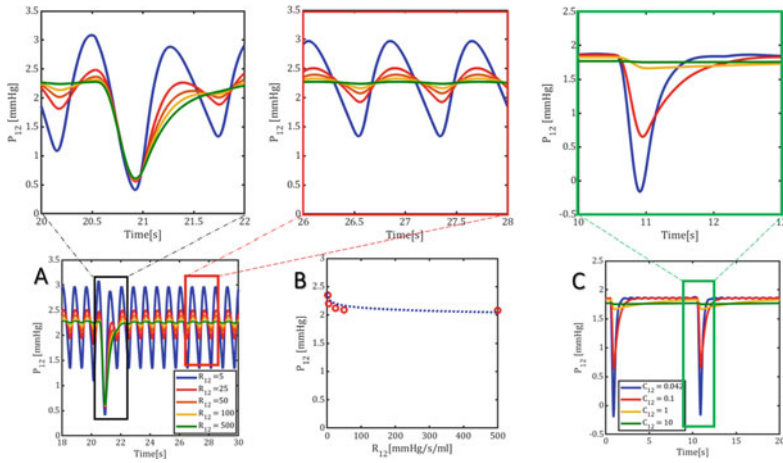


Fig. 3 Filtration process from the capillaries and extravascular resistance do not result in a negative tissue pressure in the interstitial space. **(A)** Interstitial pressure (P_{12}) is plotted over a duration of 20s for five different values of the resistance R_{12} of the interstitial space, namely 5, 25, 100, 250 , 500 mmHg s cm^{-3} . **(B)** The average pressures in the interstitial space are plotted for the different values of R_{12} . An asymptotic trend towards a positive plateau can be recognized, which leads to the conclusion that the filtration process from the capillaries and the subsequent pressure drop in the interstitium will not result in a subatmospheric pressure values in the interstitial space. **(C)** Time profile of the interstitial pressure for three different values of the interstitial compliance C_{12} . Also in this case the mean value of the interstitial pressure remains positive

Table 2 Average values ($\overline{P_{12}}$) and amplitudes ($A_{P_{12}}$) of the interstitial pressure for the different values of R_{12} and C_{12}

R_{12} [mmHg s cm^{-3}]	$\overline{P_{12}}$ [mmHg]	$A_{P_{12}}$ [mmHg]	C_{12} [mmHg cm^{-3}]	$\overline{P_{12}}$ [mmHg]	$A_{P_{12}}$ [mmHg]
$R_{12} = 500$	2.08	2.80	$C_{12} = 10$	1.7612	0.017
$R_{12} = 100$	2.09	2.88	$C_{12} = 1$	1.7719	0.1673
$R_{12} = 50$	2.10	2.96	$C_{12} = 0.1$	1.7695	1.2305
$R_{12} = 25$	2.12	3.10	$C_{12} = 0.042$	1.7685	2.0655
$R_{12} = 5$	2.19	3.64			

time oscillations in P_{12} , see Fig. 3a. In particular, when $R_{12} = 500 \text{ mmHg s cm}^{-3}$, the pressure oscillations associated with the cardiac frequency are not observable and the overall waveform amplitude is reduced by 30% compared to the case where $R_{12} = 5 \text{ mmHg s cm}^{-3}$ (see Table 2). Although a finite number of effective interstitial resistances have been simulated, the trend in the simulation results observed in Fig. 3b suggests that the mean interstitial pressure will remain positive regardless of the value of the interstitial resistance. Fig. 3c illustrates the time profile of the interstitial pressure for different values of the compliance C_{12} . Variations in the interstitial compliance impact the amplitude of the pressure waveform in the interstitial space

but have a minimal influence on the average pressure value, which remains positive in all the simulated scenarios, presenting a maximum variation of less than 1% (see Table 2). However, it is worth noting that short negative pressure peaks are attained when C_{12} is very small, even though the average value of the pressure remains positive. Furthermore, increasing C_{12} leads to a marked dampening of the pressure oscillation associated with the lymphangion contractions.

These results infer that the interplay between the pumping action of the lymphangions and the gradients in hydrostatic pressure between the vascular and interstitial spaces promote lymph formation. Of particular importance, lymph formation occurs in the absence of subatmospheric pressure. Thus, we hypothesize that another mechanism is necessary to obtain subatmospheric interstitial pressures. To test this hypothesis, we consider Model 2, where the additional mechanism is represented by a current source. Two simulation scenarios have been considered by varying the diastolic time of the lymphangion contraction.

Notably, Model 2 yields results that are consistent with the physiological expectations. Figure 4a reports the average pressures in the capillaries (P_4), interstitial space (P_{12}), lymphangions (P_{14}) and veins (P_6) simulated via Model 2. Now, the average pressure in the interstitial space is slightly subatmospheric. Figure 4b also confirms that the time profile of the interstitial pressure is influenced by the timing and amplitude of the lymphangion relaxation. Specifically, a smaller subatmospheric value of the interstitial pressure is observed for a longer relaxation time in the lymphangions.

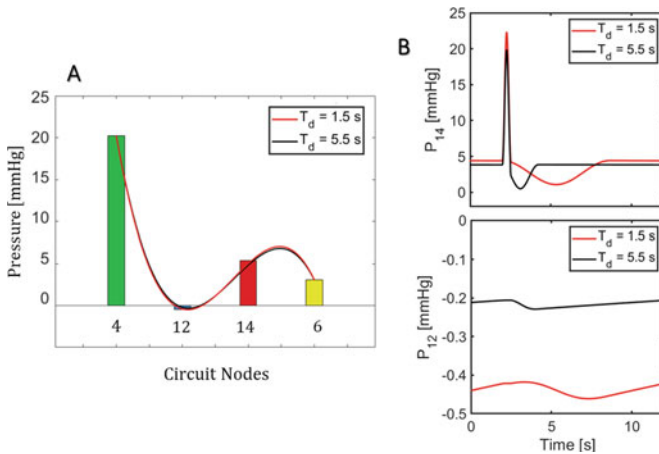


Fig. 4 (A) Average pressures in the capillaries (P_4), interstitial space (P_{12}), lymphangions (P_{14}) and veins (P_6) simulated via Model 2 for two different lymphangion diastolic times, specifically $T_d = 1$ s (red curve) and $T_d = 5.5$ s (black curve). In both configurations, the pressure decays from the value of approximately 20 mmHg in the capillaries towards a slightly negative value in the interstitial space (0.42 mmHg and 0.21 mmHg, respectively), as physiologically expected. Then, the mean pressure rises up as a result of the lymphangion pumping, leading to a value of 5.01 and 4.85 mmHg respectively. Pressure then falls towards the pressure value observed in the venous compartment of the cardiovascular circulation (approximately 3 mmHg in both configuration). (B) Pressure waveform in the lymphangions (P_{14}) (top) and in the interstitial space (P_{12}) (bottom)

4 Conclusions

The mechanisms establishing subatmospheric pressure values in the interstitial space have been debated since the publication of the work by Guyton et al. in 1963. Despite the conjectures of the existence of a force withdrawing fluid [15], also defined as a ‘lymphatic pump’ in the review by Taylor et al. [7], a quantitative demonstration of this conjecture has not been provided. In addition, the majority of the computational literature has focused on the cardiovascular system, leaving the lymphatic system and its connection with the interstitial space widely understudied, despite playing a fundamental role in maintaining fluid homeostasis in health and disease. To address these issues, this work presents a closed-loop lumped parameter model that comprises four interconnected human compartments: heart, systemic circulation, lymphatic system, and pulmonary circulation. Two versions of the model are compared. They are based on the same description for the cardiovascular system and lymphatic return but differ in the description of the passage of fluid from the interstitial space to the lymphatic system (lymph formation). The proposed formulation has the novelty of embedding both major routes of circulation, cardiovascular and lymphatic, as well as the interstitial spaces within an integrated and systemic computational environment. Model simulations suggest that filtration and pumping processes may not be the sole mechanism responsible for subatmospheric interstitial pressures. Additionally, the subatmospheric pressure in the interstitium was observed to be an intrinsic component of lymphatic function and should not be regarded as an obstacle that the lymphatic system must overcome. Because interstitial pressure and volume have crucial roles in maintaining fluid homeostasis in the human body, results from the present study suggest that the functional relationship between interstitial spaces and the lymphatic system is a necessary factor that should be considered in pathologies associated with impaired lymphatic activity, such as lymphedema and cancer metastasis. This preliminary study also emphasized the paucity of experimental studies on the coupling between cardiovascular and lymphatic functions. As new data will become available, it will be possible to reassess hypotheses and conclusions presented in this study and deepen our understanding of these complex, coupled systems.

References

1. Benoit, J.N., Zawieja, D.C., Goodman, A.H., et al.: Characterization of intact mesenteric lymphatic pump and its responsiveness to acute edemagenic stress. *Am. J. Physiol. Circ. Physiol.* **257**, H2059–H2069 (1989)
2. Papp, M., Makara, G.B., Hajtman, B.: The resistance of in situ perfused lymph trunks and lymph nodes to flow. *Cell. Mol. Life Sci.* **27**, 391–392 (1971)
3. Drake, R.E., Allen, S.J., Williams, J.P., et al.: Lymph flow from edematous dog lungs. *J. Appl. Physiol.* **62**, 2416–2420 (1987)

4. Unno, N., Nishiyama, M., Suzuki, M., et al.: A novel method of measuring human lymphatic pumping using indocyanine green fluorescence lymphography. *J. Vasc. Surg.* **52**, 946–952 (2010)
5. Olszewski, W.L., Engeset, A.: Intrinsic contractility of prenodal lymph vessels and lymph flow in human leg. *Am. J. Physiol. Circ. Physiol.* **12**, 81–84 (1980)
6. Himeno, Y., Ikebuchi, M., Maeda, A., et al.: Mechanisms underlying the volume regulation of interstitial fluid by capillaries: a simulation study. *Integr. Medicine Res.* **5**, 1121 (2016)
7. Taylor, A.E., Gibson, W.H., Granger, H.J., et al.: Review in lymphology—the interaction between intracapillary and tissue force in the overall regulation of interstitial fluid volume. *Lymphology* **6**, 192–208 (1973)
8. Venugopal, A.M., Stewart, R.H., Laine, G.A., et al.: Lymphangion coordination minimally affects mean flow in lymphatic vessels. *Am. J. Physiol. Circ. Physiol.* **293**, H1183–H1189 (2007)
9. Venugopal, A.M., Stewart, R.H., Laine, G.A., et al.: Edemagenic gain and interstitial fluid volume regulation. *Am. J. Physiol. Integr. Comp. Physiol.* **294**, R651–R659 (2008)
10. Mendoza, E., Schmid-Schnbein, G.W.: A model for mechanics of primary lymphatic valves. *Transaction—Am. Soc. Mech. Eng. J. Biomech. Eng.* **125**, 407–414 (2003)
11. Reddy, N.P., Patel, K.: A mathematical model of flow through the terminal lymphatics. *Med. Eng. Phys.* **17**, 134–140 (1995)
12. Bertram, C.D., Macaskill, C., Moore, J.E.: Simulation of a chain of collapsible contracting lymphangions with progressive valve closure. *J. Biomech. Eng.* **133**, 011008 (2011)
13. Guidoboni, G., Sala, L., Enayati, M. et al.: Cardiovascular function and ballistocardiogram: a relationship interpreted via mathematical modeling. *IEEE Transactions on Biomed. Eng.* **66**, 2906–2917 (2019)
14. Guyton, A.C., Frank, M., Abernathy, B.: A concept of negative interstitial pressure based on pressures in implanted perforated capsules. *Circ. Res.* **12**, 399–414 (1963)
15. Levick, J.R.: An investigation into the validity of subatmospheric pressure recordings from synovial fluid and their dependence on joint angle. *The. J. Physiol.* **289**, 55–67 (1979)
16. Jamalian, S., Jafarnejad, M., Zawieja, S.D., et al.: Demonstration and analysis of the suction effect for pumping lymph from tissue beds at subatmospheric pressure. *Sci. Reports* **7**, 12080 (2017)
17. Avanzolini, G., Barbini, P., Cappello, A. et al.: Time-varying mechanical properties of the left ventricle—a computer simulation. *IEEE Transactions on Biomed. Eng. BME* **32**(10), 756–763 (1985)
18. Braakman, R., Sipkema, P., Westerhof, N.: A dynamic nonlinear lumped parameter model for skeletal muscle circulation. *Annals Biomed. Eng.* **17**, 593–616 (1989)
19. Muller, L.O., Toro, E.F.: A global multiscale mathematical model for the human circulation with emphasis on the venous system. *Int. J. for Numer. Methods Biomed. Eng.* **30**, 681–725 (2014)
20. Aukland, K., Reed, R.K.: Interstitial-lymphatic mechanisms in the control of extracellular fluid volume. *Physiol. Rev* **73**, 1–78 (1993)
21. Wiig, H., Swartz, M.A.: Interstitial fluid and lymph formation and transport: physiological regulation and roles in inflammation and cancer. *Physiol. Rev.* **92**, 1005–1060 (2012)
22. Avanzolini, G., Barbini, P., Cappello, A., et al.: CADCS simulation of the closed-loop cardiovascular system. *Int. J. Biomed. Comput.* **22**, 39–49 (1988)
23. Scallan, J.P., Wolpers, J.H., Muthuchamy, M., et al.: Independent and interactive effects of preload and afterload on the pump function of the isolated lymphangion. *Am. J. Physiol. Hear. Circ. Res.* **303**, H809–H823 (2012)
24. Margaris, K.N., Black, R.A.: Modelling the lymphatic system: challenges and opportunities. *J. Royal Soc.* **9**, 601–612 (2012)
25. Quick, C.M., Venugopal, A.M., Dongaonkar, R.M. et al.: First-order approximation for the pressure-flow relationship of spontaneously contracting lymphangions. *Am. J. Physiol. Circ. Physiol.* **294**, H2144–H2149 (2008)

26. Sloas, D.C., Stewart, S.A., Sweat, R.S., et al.: Estimation of the pressure drop required for lymph flow through initial lymphatic networks. *Lymphatic Res. Biol.* **14**(2), 62–69 (2016)
27. Kunert, C., Baish, J.W., Liao, S., et al.: Mechanobiological oscillators control lymph flow. *Proc. Natl. Acad. Sci.* **112**(35), 10938–10943 (2015)
28. Roose, T., Swartz, M.A.: Multiscale modeling of lymphatic drainage from tissues using homogenization theory. *J. Biomech.* **45**(1), 107–115 (2012)
29. MathWorks Inc. MATLAB: application program interface guide. Mathworks (5) (1996)
30. Fritzson, P.: Principles of object-oriented modeling and simulation with Modelica 2.1. Springer Science & Business Media (2010)
31. Scallan, J.P., Huxley, V.H., Korthuis, R.J.: Capillary fluid exchange: regulation, functions and pathology. Springer Science & Business Media (1) (2010)
32. Formaggia, L., Quarteroni, A., Veneziani, A.: Cardiovascular mathematics: modeling and simulation of the circulatory system. Morgan and Claypool Life Science Publisher, Colloquium Lectures on Integrated Systems Physiology-From Molecules to Function (2010)
33. Guyton, A.C., Hall, J.E.: The microcirculation and the lymphatic system: capillary fluid exchange. Interstitial Fluid, and Lymph Flow. *Textbook of Medical Physiology* 11th ed. (2006)
34. Mortimer, P.S., Rockson, S.G.: New developments in clinical aspects of lymphatic disease. *J. Clin. Invest.* **124**(3), 915–921 (2014)

Age-Structured Epidemic with Adaptive Vaccination Strategy: Scalar-Renewal Equation Approach



Aubain Nzokem and Neal Madras

Abstract We use analytical and numerical methods to investigate an adaptive vaccination strategy's effects on the infectious disease dynamics in a demographically open population. The methodology and key assumptions are based on Breda et al. (2012). We show that the endemic force of infection in the demographically open population can be reduced significantly by two factors: the vaccine effectiveness and the vaccination rate. The impact of these factors can transform an endemic steady state into a disease free state.

Keywords Force of infection · Scalar-renewal equation · Per capita death rate · Adaptive vaccination strategy

1 Introduction

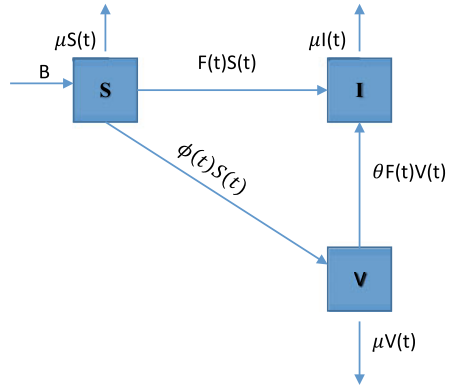
The 1927 paper of Kermack and McKendrick [5] is one of the fundamental contributions to the mathematical theory of epidemic modelling. The paper provides the condition of outbreak and the final size equation in a closed population setting. One of the key features of [5] was to introduce an age of infection model. In such a model, the general infectivity function ($A(\tau)$) of an individual is considered and depends on the time (τ) elapsed since the infection took place. Kermack and McKendrick's framework encompasses a wide family of epidemic models; Breda et al. [2] have illustrated the generalisation by providing the following age infectivity function for SIR and SEIR models.

A. Nzokem (✉) · N. Madras
Department of Mathematics & Statistics, York University, 4700 Keele St.,
Toronto, ON M2R 3R4, Canada
e-mail: aubain14@yorku.ca

N. Madras
e-mail: madras@yorku.ca

© Springer Nature Switzerland AG 2021
D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_54

Fig. 1 The transfer diagram of the model: the force of infection function $t \mapsto F(t)$; the rate of vaccination function $t \mapsto \phi(t)$; θ (vaccine parameter ($0 < \theta < 1$)); I (infected population); B (constant birth rate); μ (constant per capita death rate)



$$\begin{aligned}
 A(\tau) &= \beta e^{-\alpha\tau} && \iff SIR \\
 A(\tau) &= \beta \frac{\gamma}{\gamma - \alpha} (e^{-\alpha\tau} - e^{-\gamma\tau}) && \iff SEIR \tag{1}
 \end{aligned}$$

The 2012 paper of Breda et al. *On the formulation of epidemic models (an appraisal of Kermack and McKendrick)* [2], revised the original Kermack and McKendrick paper [5] and produced the same results, but the method used was different. In fact, Breda et al. [2] considered the unknown force of infection as a result of the nonlinear scalar-renewal equation; and they analyzed the force of infection at endemic equilibrium. For related work of interest, see [1, 3, 4, 6, 7].

In the current paper, we investigate the effects of an adaptive vaccination strategy on the dynamics of infectious diseases in a demographically open population. In contrast to standardized childhood vaccinations, we consider a situation in which an individual may get vaccinated at any age in response to rising prevalence. The decision for an individual to get vaccinated could be a response to the increasing perceived threat of infection, or perhaps due to the greater availability of the vaccine as production is ramped up in response to the epidemic. Our model, while fairly general, makes two restrictive assumptions: first, that the disease does not affect death rates; and second, that at any time the rate of vaccination depends on the current force of infection (i.e., the time lag is negligible).

The methodology and key assumptions are based on [2]. Individuals’ lifetimes have an arbitrary probability distribution. The epidemic model and the vaccination process is illustrated by Fig. 1 below. The susceptible population is divided into non-vaccinated susceptible population (S) and vaccinated susceptible population (V). The infection leads to permanent immunity (no re-infection), although infectivity varies according to the function $A(\tau)$ mentioned earlier. Infection status does not affect the time of death of an individual.

Our main assumptions are that the force of infection among the vaccinated susceptibles varies proportionally with the force of infection among the unvaccinated susceptibles; and the rate of vaccination is proportional to the force of infection in

the unvaccinated susceptibles. The natural feature of an adaptive vaccination policy is that the rate of vaccination should increase when the force of infection increases and decrease when the force of infection decreases. Direct proportionality represents a simple case with the advantage of being nicely tractable.

The rest of this paper will analyse an age-structured epidemic model with adaptive vaccination. A complete analysis, including a study of an analogous model without age structure, appears in [8, 9].

2 Age-Structured Epidemic Model

We consider the situation where, at the population level, new susceptibles are born at a constant rate B . We also consider a general survival function $\mathcal{F}(a)$, which describes the probability that a newborn individual lives at least until age a . As is well known, $\int_0^\infty \mathcal{F}(a) da$ equals the average lifetime, which is the reciprocal of the effective per capita death rate μ when $\mathcal{F}(a) = e^{-\mu a}$.

If at time t a susceptible has age a , then at time $t - a + \sigma$ the susceptible had age σ ($0 < \sigma \leq a$). For a small value h , taking the survival function $\mathcal{F}(a)$ into account, the time evolution of the unvaccinated susceptibles $S(t, a)$ at time t and at age a satisfies the following equation:

$$\frac{S(t - a + \sigma + h, \sigma + h)}{\mathcal{F}(\sigma + h)} = \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} (1 - F(t - a + \sigma)h - \phi(t - a + \sigma)h + o(h^2)) \tag{2}$$

By re-arranging and using the limit as h converges to 0, we will have the following derivative:

$$\frac{d\left(\frac{S(t-a+\sigma,\sigma)}{\mathcal{F}(\sigma)}\right)}{d\sigma} = -F(t - a + \sigma) \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} - \phi(t - a + \sigma) \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} \tag{3}$$

A similar approach can be used to derive the differential equation for vaccinated susceptibles $V(t, a)$ at time t and at age a . The dynamics of both classes of susceptibles can be described by the system of differential Eq. (4).

$$\begin{aligned} \frac{d\left(\frac{S(t-a+\sigma,\sigma)}{\mathcal{F}(\sigma)}\right)}{d\sigma} &= -F(t - a + \sigma) \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} - \phi(t - a + \sigma) \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} \\ \frac{d\left(\frac{V(t-a+\sigma,\sigma)}{\mathcal{F}(\sigma)}\right)}{d\sigma} &= -\theta F(t - a + \sigma) \frac{V(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} + \phi(t - a + \sigma) \frac{S(t - a + \sigma, \sigma)}{\mathcal{F}(\sigma)} \end{aligned} \tag{4}$$

The force of infection, of course, depends heavily on the size of the infectious population. At time t , considering individuals who were infected at time $t - \tau$ at age a , the contribution to the force of infection is the product of $(F(t - \tau)S(t - \tau, a) + \theta F(t - \tau)V(t - \tau, a))A(\tau)$ infectious individuals and a demographic factor $\frac{\mathcal{F}(a+\tau)}{\mathcal{F}(a)}$, which is the proportion of infectious individuals of age a who survive to age $a + \tau$.

By summing all the contributions with respect to to the elapsed time τ and with respect to the age a , we get the following scalar-renewal equation.

$$\begin{aligned}
 F(t) &= \int_0^\infty \int_0^\infty (F(t - \tau)S(t - \tau, a) + \theta F(t - \tau)V(t - \tau, a))A(\tau) \frac{\mathcal{F}(a + \tau)}{\mathcal{F}(a)} d\tau da \\
 &= \int_0^\infty F(t - \tau) \int_0^\infty (S(t - \tau, a) + \theta V(t - \tau, a)) \frac{\mathcal{F}(a + \tau)}{\mathcal{F}(a)} A(\tau) d\tau da \quad (5)
 \end{aligned}$$

Generally, the integral $\int_0^a \phi(t - a + \sigma)e^{-\int_0^\sigma ((1-\theta)F + \phi)(t-a+\tau)d\tau} d\sigma$ is difficult to evaluate because the rate of vaccination function ($\phi(t)$) is unknown. As we previously assumed, there is a linear relationship between the vaccination rate and the force of infection. We have $\phi(t - a + \sigma) = pF(t - a + \sigma)$ where p is the vaccination rate parameter. The solution of the system of differential Eq. (4) becomes:

$$\begin{aligned}
 S(t, a) &= B \mathcal{F}(a)e^{-(1+p) \int_0^a F(t-a+\sigma)d\sigma} \\
 V(t, a) &= B \mathcal{F}(a) \frac{P}{1 + p - \theta} \left\{ e^{-\theta \int_0^a F(t-a+\sigma)d\sigma} - e^{-(1+p) \int_0^a F(t-a+\sigma)d\sigma} \right\} \quad (6)
 \end{aligned}$$

2.1 Equilibrium Equation of the Endemic Steady State

The solution (6) can be substituted in the renewal Eq. (5). When the parameters permit an endemic steady state, the force of infection ($F(t)$) converges to a positive constant F . When t goes to $+\infty$, the renewal equation can be rearranged, leading to the following equilibrium equation of the endemic steady state.

$$1 = \frac{B}{1 + p - \theta} \int_0^\infty \int_0^\infty (p\theta e^{-\theta aF} + (1 + p)(1 - \theta)e^{-(1+p)aF}) \mathcal{F}(a + \tau)A(\tau) d\tau da \quad (7)$$

In order to study the properties of the equilibrium equation, we consider the following function.

$$f(x) = \frac{B}{1 + p - \theta} \int_0^\infty \int_0^\infty (p\theta e^{-\theta ax} + (1 + p)(1 - \theta)e^{-(1+p)ax}) \mathcal{F}(a + \tau)A(\tau) d\tau da \quad (8)$$

It is obvious that $f(x)$ is a decreasing function and $f(0) = B \int_0^\infty \int_0^\infty \mathcal{F}(a + \tau)A(\tau) d\tau da$. The condition $f(0) > 1$ is sufficient to guarantee the existence of F with $F > 0$. Furthermore, $f(0)$ depends only on the constant birth rate B , the survival function $\mathcal{F}(\cdot)$ and the expected contribution to the force of infection $A(\cdot)$. The vaccination parameters θ and p cancel out in $f(0)$. The quantity $f(0)$ is the basic reproduction number:

$$R_0 = f(0) = B \int_0^\infty \int_0^\infty \mathcal{F}(a + \tau)A(\tau)d\tau da \quad (9)$$

2.2 General Case

We shall treat the quantity F in Eq. (7) as an implicit function of two variables (θ, p) . Writing this explicitly in the equation, we have

$$1 = \frac{B}{1+p-\theta} \int_0^\infty \int_0^\infty (p\theta e^{-\theta a F(\theta,p)} + (1+p)(1-\theta)e^{-(1+p)aF(\theta,p)}) \times \mathcal{F}(a+\tau)A(\tau) d\tau da. \tag{10}$$

Case 1: ineffective vaccine ($\theta = 1$)

The Eq. (10) becomes:

$$1 = B \int_0^\infty \int_0^\infty e^{-aF(1,p)} \mathcal{F}(a+\tau)A(\tau) d\tau da \tag{11}$$

We have $F(1, p) = F^*$, where F^* comes from the unvaccinated case studied by [2]. Therefore, the endemic force of infection is the same as the endemic force of infection without vaccination.

Case 2: 100% effective vaccine ($\theta = 0$)

The Eq. (10) becomes:

$$1 = B \int_0^\infty \int_0^\infty e^{-(1+p)aF(0,p)} \mathcal{F}(a+\tau)A(\tau) d\tau da \tag{12}$$

As previously, we have $(1+p)F(0, p) = F^*$ for $p > 0$. we can deduce that $F(0, p) = \frac{F^*}{p+1}$, which depends on the factor $\frac{1}{p+1}$. $\lim_{p \rightarrow \infty} F(0, p) = 0$, which corresponds to the disease free steady state.

Case 3: $p \rightarrow \infty$ and $\theta \neq 0$

The Eq. (10) becomes:

$$1 = B\theta \int_0^\infty \int_0^\infty e^{-\theta a F(\theta,+\infty)} \mathcal{F}(a+\tau)A(\tau) d\tau da \tag{13}$$

The solution of the Eq. (13) is

$$F(\theta, +\infty) = \begin{cases} 0 & \text{if } \theta \leq \frac{1}{f(0)} \\ x(\theta) & \text{if } \frac{1}{f(0)} < \theta < 1 \\ F^* & \text{if } \theta = 1 \end{cases}$$

where $x(\theta)$ is an increasing function with $0 < x(\theta) < F^*$.

2.3 Special Case of Natural Constant Per-Capita Mortality Rate

It is of interest to examine the results of the previous subsection in a case where $F(\theta, p)$ can be solved exactly. Accordingly, we shall assume in this subsection that all individuals have a survival function $\mathcal{F}(a) = e^{-\mu a}$, which describes a constant per-capita mortality rate μ . By applying the survival function, the following basic reproduction number is derived from Eq. (9):

$$\begin{aligned}
 f(0) = R_0 &= B \int_0^\infty \int_0^\infty e^{-\mu(a+\tau)} A(\tau) d\tau da \\
 &= \frac{1}{\mu} B \int_0^\infty e^{-z\tau} A(\tau) d\tau
 \end{aligned}
 \tag{14}$$

The reproduction number is the same as that found by [2] for a constant per-capita mortality rate μ . The equilibrium Eq. (10) for endemic steady state becomes the second degree equation

$$\frac{\theta(1+p)}{\mu f(0)} F(\theta, p)^2 + \left(\frac{1+p+\theta}{f(0)} - \theta(1+p) \right) F(\theta, p) + \frac{\mu}{f(0)} (1-f(0)) = 0.
 \tag{15}$$

We define

$$a = \frac{\theta(1+p)}{\mu f(0)}, \quad b = \frac{1+p+\theta}{f(0)} - \theta(1+p), \quad c = \frac{\mu}{f(0)} (1-f(0)).$$

In endemic steady state, $f(0) > 1$. We have $c = \frac{\mu}{f(0)} (1-f(0)) < 0$ and $b^2 - 4ac > 0$. The solution of the Eq. (15) becomes

$$\begin{aligned}
 F(\theta, p) &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \\
 &= \frac{\mu}{2} \left\{ f(0) - \frac{1+p+\theta}{\theta(1+p)} + \sqrt{\left(\frac{1+p+\theta}{\theta(1+p)} - f(0) \right)^2 + 4 \frac{f(0)-1}{\theta(1+p)}} \right\}
 \end{aligned}
 \tag{16}$$

With our vaccination parameters, the expression for the endemic force of infection becomes more complex. By comparison, in [2], with a constant per-capita mortality rate μ and no vaccination, the authors find the endemic force of infection $F^* = \mu(f(0) - 1)$.

Case 1: Ineffective vaccine ($\theta = 1$)

The endemic force of infection becomes

$$\begin{aligned} F(1, p) &= \lim_{\theta \rightarrow 1} F(\theta, p) \\ &= \frac{\mu}{2} \left\{ f(0) - \frac{2+p}{1+p} + \sqrt{\left(\frac{2+p}{1+p} - f(0)\right)^2 + 4\frac{f(0)-1}{1+p}} \right\} \\ &= \mu(f(0) - 1) \\ &= F^* . \end{aligned}$$

In the case of 100% ineffective vaccine, the endemic force of infection is the same as the endemic force of infection without vaccination [2].

Case 2: 100% effective vaccine ($\theta = 0$)

The endemic force of infection becomes

$$\begin{aligned} \lim_{\theta \rightarrow 0} F(\theta, p) &= \lim_{\theta \rightarrow 0} \frac{\mu}{2} \left[f(0) - \frac{1+p+\theta}{\theta(1+p)} + \sqrt{\left(\frac{1+p+\theta}{\theta(1+p)} - f(0)\right)^2 + 4\frac{f(0)-1}{\theta(1+p)}} \right] \\ &= \mu \frac{(f(0) - 1)}{1+p} \\ &= \frac{F^*}{1+p} . \end{aligned}$$

The quantity $F^* = \mu(f(0) - 1)$ is the endemic force of infection without vaccination from [2]. Taking into account the vaccination process, we have $\lim_{\theta \rightarrow 0} F(\theta, p) = \frac{F^*}{p+1}$, which depends on the factor $\frac{1}{p+1}$ with the vaccination rate parameter p . We have $\lim_{p \rightarrow \infty} \{\lim_{\theta \rightarrow 0} F(\theta, p)\} = 0$, which corresponds to the disease free steady state.

Case 3: $p \rightarrow \infty$ and $\theta \neq 0$

Here the vaccine is not 100% effective. By increasing the vaccination rate parameter (p), the expression in (16) becomes

$$\lim_{p \rightarrow \infty} F(\theta, p) = \begin{cases} 0 & \text{if } \theta \leq \frac{1}{f(0)} , \\ \mu(f(0) - \frac{1}{\theta}) & \text{if } \frac{1}{f(0)} < \theta \leq 1 . \end{cases}$$

With a sufficiently high vaccination rate parameter (p), the disease free steady state can still be reached if the vaccine parameter (θ) is below a threshold ($\theta \leq \frac{1}{f(0)}$).

Case 4: $0 < \theta < 1$ and $p > 0$

The endemic force of infection in the expression (16) was simulated as a function of vaccination rate parameter (p) and vaccine parameter (θ). Figure 2 provides a sum-

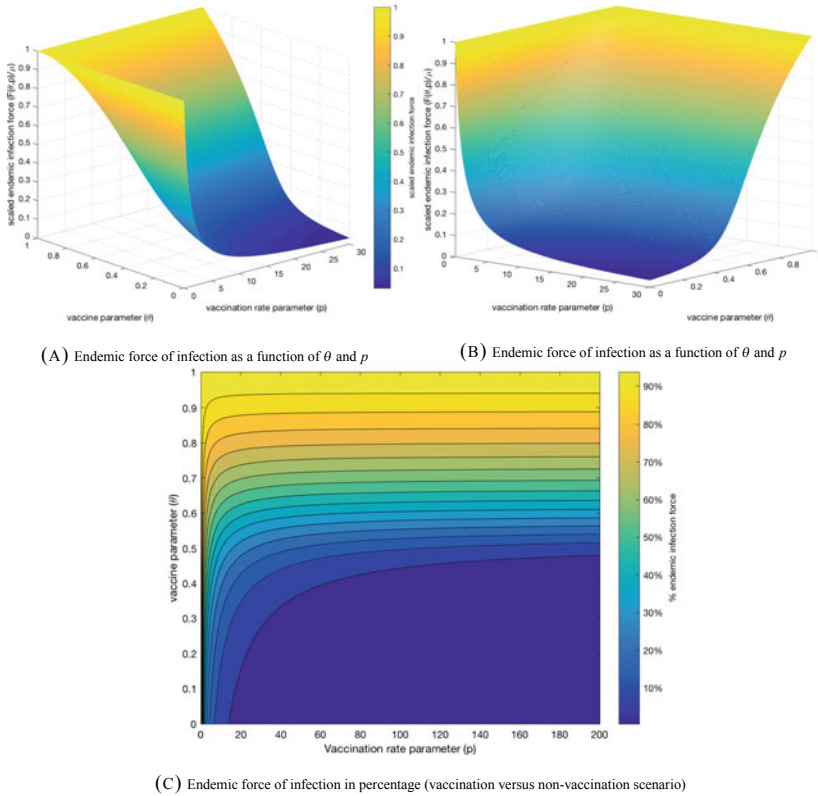


Fig. 2 Impact of adaptive vaccination strategy on the endemic force of infection ($R_0 = f(0) = 2$)

mary of the findings. We supposed the reproduction number is $R_0 = 2$. As illustrated by the yellow area in Fig. 2, the scaled endemic infection force ($\frac{F^*}{\mu}$) remains almost constant; whereas in the blue area, the reduction of the endemic force of infection is significant. Compare to the unvaccinated scenario, the endemic force of infection is almost 0.

3 Conclusion

The relation between the force of infection of the disease spreading, the vaccine effectiveness, and the vaccination rate is at the center of the article. The investigation focuses on a demographically open population using the renewal equation on the force of infection. The findings show that the endemic force of infection in a demographically open population can be reduced significantly by a good combination of the vaccine effectiveness and the vaccination rate. In fact, for a given vaccination rate

parameter (p), it is shown that the endemic force of infection can remain unchanged if the vaccine is ineffective; whereas the endemic force of infection converges to a disease free steady state when the vaccine is 100% effective. It is also shown that a sufficiently high vaccination rate parameter can transform an endemic steady state into a disease free state when the vaccine is adequately effective in the sense that $\theta \leq \frac{1}{f(0)}$. One of the limitations of the study is the linearity assumptions on the rate of vaccination function and the force of infection within the vaccinated susceptibles.

Acknowledgements We would like to thank Prof. Jianhong Wu and Mahnaz Alavinejad for informative discussions, and the referees for helpful comments.

References

1. Andersson, H., Britton, T.: Stochastic epidemic models and their statistical analysis. Springer-Verlag, New York (2000)
2. Breda, D., Diekmann, O., de Graaf, W.F., Pugliese, A., Vermiglio, R.: On the formulation of epidemic models (an appraisal of Kermack and McKendrick). *J. Biol. Dynam.* **6**, 103–117 (2012)
3. Diekmann, O., Heesterbeek, H., Britton, T.: Mathematical tools for understanding infectious disease dynamics. Princeton University Press, Princeton (2013)
4. Huppert, A., Katriel, G.: Mathematical modelling and prediction in infectious disease epidemiology. *Clin. Microbiol. Infect.* **19**, 999–1005 (2013)
5. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond.* **115**, 700–721 (1927)
6. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics II. The problem of endemicity, *Proc. R. Soc. Lond.* **138**, 55–83 (1932)
7. Meehan, M.T., Cocks, D.G., Müller, J., McBryde, E.S.: Global stability properties of a class of renewal epidemic models. *J. Math. Biol.* **78**, 1713–1725 (2019)
8. Nzokem, A., Madras, N.: Epidemic dynamics and adaptive vaccination strategy: renewal equation approach. *Bull. Math. Biol.* **82**(9), 122 (2020)
9. Nzokem, A.: Stochastic and renewal methods applied to epidemic models. PhD thesis, York University, YorkSpace Institutional Repository (2020)

On the Modeling of Drug Delivery to Solid Tumors; Computational Viewpoint



Mohsen Rezaeian , Madjid Soltani , and Farshad Moradi Kashkooli 

Abstract Drug distribution in a solid tumor is important in the evaluation of cancer treatment efficacy. In the present study, a comprehensive multi-scale mathematical model employed to calculate the interstitial fluid pressure (IFP) and drug distribution in interstitial space. Two different zones including necrotic core and semi-necrotic zone are considered for a tumor surrounded by normal tissue. The results indicate that drug concentration has its maximum value in the semi-necrotic region and it starts declining steeply in the necrotic area. Different sizes of the necrotic core are considered by introducing the ratio $R_n = r_n/r_t$, where r_t and r_n are tumor and necrotic core radius, respectively. Generally, increasing R_n leads to a decrease in IFP. The decrease in IFP is more significant at larger R_n values. In addition, IFP is more sensitive to R_n values in smaller tumors. While it is expected that with a decreased IFP, drug transport to the tumor will be facilitated due to the reduced outward convection flow in the tumor interstitium, the mean drug concentration in tumor decreases with increasing R_n . These findings show that such a mathematical model is a powerful tool and provides more insight into the drug's transport and delivery to solid tumors.

Keywords Drug delivery · Solid tumor · Porous media · Necrotic core · Drug concentration

M. Rezaeian · M. Soltani · F. Moradi Kashkooli
Department of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran
e-mail: mohsenrezaeian@gmail.com

M. Soltani (✉)
Centre for Biotechnology and Bioengineering (CBB), University of Waterloo, Waterloo, ON,
Canada
e-mail: msoltani@uwaterloo.ca

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON,
Canada

F. Moradi Kashkooli
Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_55

601

1 Introduction

Although many anticancer drugs have been developed in recent years, they are often ineffective. The ineffectiveness of cancer drugs is related to the complex microenvironment of solid tumors and drug properties. Therefore, better understanding of the complex tumor microenvironment and mechanisms of drug delivery to solid tumors is crucial in designing an effective treatment strategy. Since many factors are involved in drug delivery, it is almost impossible to experimentally and economically investigate all factors thorough clinical and preclinical studies. In contrast, mathematical modeling can play an effective role. The aim of mathematical modeling and simulation is to better understand the behaviors of the tumor for ultimately improving the treatment outcome. Thus using mathematical modeling, the aim of this work is to comprehensively investigate variables that affect drug delivery to solid tumors.

In solid tumors, physiological barriers are the major cause of decreasing the efficacy of drug delivery [1]. These barriers have a contribution in high values of Interstitial fluid pressure (IFP) in solid tumors. High IFP value prevents the systematic drug delivery by causing an outward convection against inward drug diffusion [2, 3]. Elevated IFP produces interstitial fluid flow (IFF) outward from the tumor center and conveys tumor-produced macromolecules such as Vascular Endothelial Growth Factors (VEGFs) toward the normal tissue and also hinders drugs reaching most of the cancerous cells [4]. According to the imbalance between supply and demand of nutrients and oxygen in the fast growing tumor, most tumor nodules have a necrotic core in which there are no living cells or vascular system, therefore, there is no blood flow or cellular uptake [2]. The presence of a necrotic area in the tumor is one of the main reasons for the heterogeneous distribution of the drug in the tumor, and therefore the size of this area may be an influencing factor in the process of drug delivery to the tumor.

A number of studies have been conducted on the transport of drug molecules into solid tumors. The first formulation to calculate the concentration distribution in solid tumors is presented by Baxter et al. [5]. They employed a continuous porous media model to study the impact of various parameters such as interstitial pressure and convection, heterogeneous perfusion and lymphatics, and binding and metabolism on the drug concentration in the ECM. Also, they have studied the concentration distribution of two drugs (Fab and IgG) in an isolated circular tumor. Soltani and Chen [6–8] developed the fluid flow mathematical model and employed it to study two new parameters of critical radius of the tumor and critical radius of the necrotic zone on IFP distribution in solid tumors. Also, they used developed model for different geometries to investigate the impacts of tumor shape and size on drug delivery [9, 10]. To investigate the influences of different parameters including IFP, interstitial fluid velocity (IFV), and concentration, there exist many new efforts in the field of mathematical modeling and simulation of drug delivery to solid tumor in the literature [2, 3, 11–14]. Among these, a recent study by Steuperaert et al. [14] showed that consideration of a necrotic area in the tumor had no effect on drug distribution for intraperitoneal chemotherapy, whilst the results of Soltani and Chen [6] suggest that

the size of the necrotic area could potentially change the process of drug delivery by changing the IFP value. However, among studies to date, the effect of necrotic area on the drug distribution in solid tumors has not been fully investigated and remains unknown.

A solid tumor model can be employed to evaluate and optimize the strategies of treatment for the aim of personalized medicine. In the current study, a comprehensive approach is applied to evaluate the drug delivery to a solid tumor by considering necrotic core with variable radius within the tumor. This study proposes a modeling framework to calculate IFV, IFP, as well as the transport of drug molecules and drug concentration distribution.

2 Material and Methods

This section covers mathematical modeling approach, model geometry, boundary conditions, model parameters, and verification of the results. First, a physically relevant tumor microenvironment is modeled. Then, the continuity equation, Darcy's law, and Starling's equation with appropriate boundary conditions for normal and tumor tissues are developed to calculate the IFP and IFF. To solve these equations accurately in this complex microenvironment, the finite element method is employed with an adaptive grid generation which combines high accuracy with a high calculation speed. Moreover, by using advection–diffusion equation in tumor and normal tissue, dynamic concentration profile of drug molecules in the tumor is calculated.

2.1 Mathematical Modeling Approach

Interstitial fluid flow. Here, the tissue is considered as a porous medium. For fluid transport in tissue, Darcy's law as one of the first formulations for flow transport in a porous medium is used. In the current study, the IFV is calculated by Darcy's equation, as following [2]:

$$v_i = -\kappa \nabla P_i \quad (1)$$

While the IFP is calculated as follows:

$$-\kappa \nabla^2 P_i = \phi_V - \phi_L \quad (2)$$

where ϕ_V is the net fluid flow rate per unit volume from blood vessels into the interstitium, and ϕ_L is the net flow rate per unit volume from interstitium into the lymphatic vessels. ϕ_V and ϕ_L are obtained using the Starling's law [2, 3, 6, 8]:

$$\phi_V = \frac{L_P S}{V} (P_B - P_i - \sigma_s (\pi_B - \pi_i)) \quad (3)$$

$$\phi_L = \frac{L_{PL} S_L}{V} (P_i - P_L) \quad (4)$$

where L_P is the hydraulic conductivity of the microvascular wall, $\frac{S}{V}$ is the vascular surface area per unit volume, P_B and P_i , respectively, are the intravascular blood pressure and interstitial fluid pressure, σ_s is the average osmotic reflection coefficient for plasma proteins, π_B is the plasma osmotic pressure, and π_i is the interstitial fluid osmotic pressure. Also L_{PL} is the hydraulic conductivity of the lymphatic wall, $\frac{S_L}{V}$ is the ratio of the surface area of lymphatic vessels to the tumor tissue volume, and P_L is the hydrostatic pressure of the lymphatic. Due to the lack of an effective lymphatic system in the tumor tissue, the term ϕ_L is considered to be zero.

Solute transport. Based on the conservation laws of mass and momentum; by considering a convection–diffusion mechanism, the equation for the drug transport in ECM can be written as [2, 3, 11]:

$$\frac{\partial C}{\partial t} = -v \nabla C + D \nabla^2 C + (\Phi_V - \Phi_L) \quad (5)$$

where C is the drug concentration, v is the IFV obtained from Darcy's law, D is the coefficient of diffusion, Φ_V and Φ_L are respectively the solute transport rates per unit volume from blood vessels into the interstitium, and from the interstitium into the lymphatic vessels. These two parameters are obtained by the Kedem-Katchalsky equation [15]:

$$\Phi_B = \left(\phi_B (1 - \sigma_f) C_P + \frac{PS}{V} (C_P - C) \frac{Pe}{e^{Pe} - 1} \right) \quad (6)$$

$$\Phi_L = \phi_L C \quad (7)$$

where C_P is drug concentration in the plasma, σ_f the filtration reflection coefficient, and P is the microvessel permeability coefficient for free drug. Pe is the Peclet number that determines the convection/diffusion ratio through the capillary wall defined as:

$$Pe = \frac{\phi_B (1 - \sigma_f)}{P \frac{S}{V}} \quad (8)$$

It is worth-mentioning that the model parameters of interstitial transport and solute transport for normal and tumor tissue are the same as our group's previous work [9].

2.2 Model Geometry, Computational Domain, and Boundary Conditions

The computational domain is considered as a spherical tumor with a radius r_t and its surrounding normal tissue with a radius three times larger than the radius of the tumor. The tumor also has a necrotic core with a radius r_n . (Fig. 1). Different sizes of the necrotic core is considered in this study by introducing the ratio $R_n = r_n/r_t$. The baseline values of r_t and R_n are 1 cm and 0.5, respectively.

In the current study, a bolus injection of a chemotherapy drug is considered so that the plasma concentration decreases exponentially, as following [9]:

$$C_p = C_0 \exp(-\ln(0.5)t/\tau) \tag{9}$$

C_0 is the initial concentration and τ is the drug half-time in plasma. Concentration is then non-dimensionalized by C_0 . For the boundary between the tumor and normal tissue (inner boundary 1) and the boundary between the viable tissue and necrotic core of the tumor (inner boundary 2), continuity of the IFV, IFP, and also concentration and its flux are considered as appropriate boundary conditions; where $\Omega -$ and $\Omega +$ demonstrate the tumorous and normal tissue at the boundary. Boundary conditions of the presented investigations are outlined in Table 1. For outer boundary that the interstitial pressure is constant, the boundary condition of Dirichlet-type must be applied, and for concentration, the open boundary condition is used [3].

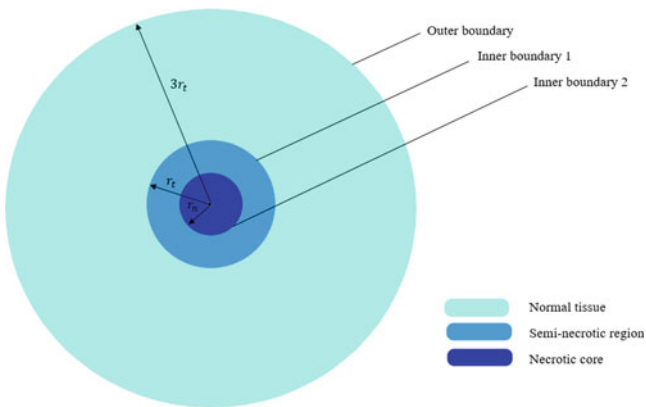


Fig. 1 Computational domain of the current study which includes tumor necrotic core, tumor semi-necrotic region, and normal tissue

Table 1 Boundary conditions employed for the present study

Region	Boundary conditions	
	Fluid flow	Concentration
Center of the tumor	$\nabla P_i = 0$	$D_{eff} \nabla C + v_i C = 0$
Inner boundaries	$-k_t \nabla P_i _{\Omega^-} = -k_n \nabla P_i _{\Omega^+}$ $P_i _{\Omega^-} = P_i _{\Omega^+}$	$(D_{eff}^t \nabla C + v_i C) _{\Omega^-} = (D_{eff}^n \nabla C + v_i C) _{\Omega^+}$ $C_i _{\Omega^-} = C_i _{\Omega^+}$
Outer boundary	$P_i = \text{Constant}$	$-n \bullet \nabla C = 0$

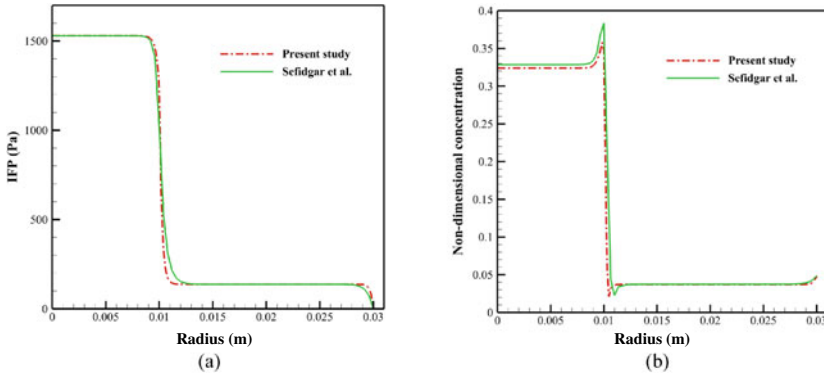


Fig. 2 Comparison of the results of the present study with the results of Sefidgar et al. [9], **a** IFP distribution, and **b** Non-dimensional concentration distribution over the non-dimensional radius

2.3 Validation

In order to verify the results of the present study, the problem addressed by Sefidgar et al. [9] has been investigated for a tumor embedded in normal tissue. As is clear in Fig. 2, the results of this study are well in correspondence with the literature [9]. In the present work, using a specific drug and parameters, we also obtain similar profiles for IFP and non-dimensional concentration with Sefidgar et al. [9]. Also, it is showing that the current models and methods are enough reliable to predict the IFP and non-dimensional concentration.

3 Results and Discussion

In the present study, a comprehensive approach for modeling the drug delivery to solid tumors considering necrotic core is employed. Convection and diffusion mechanisms of drug transport are considered in interstitial space. First, IFP and IFV are obtained based on fluid flow equations in porous media and then, drug concentrations are

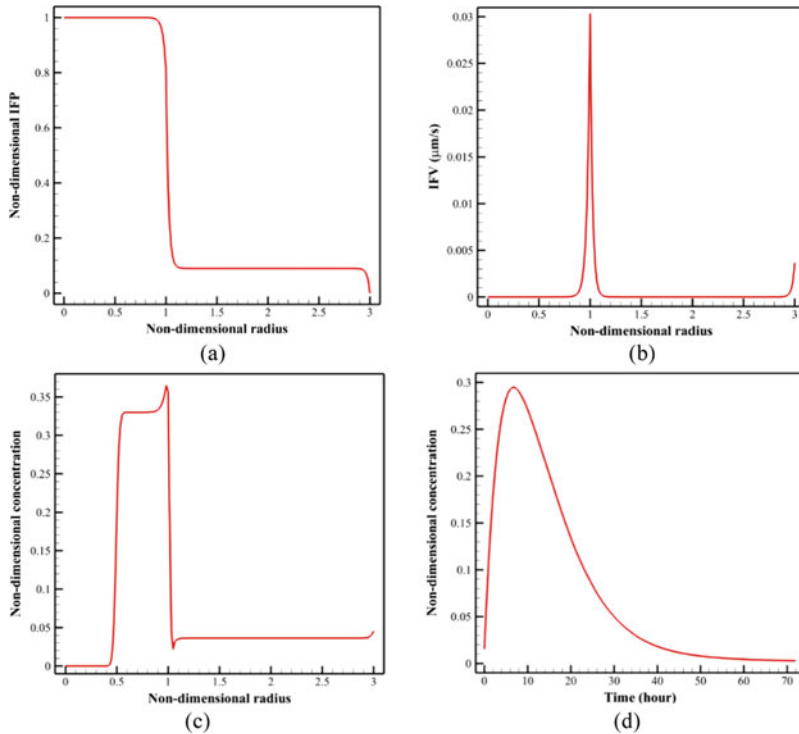


Fig. 3 The spatial distribution of **a** non-dimensional IFP, **b** IFV, and **c** non-dimensional concentration over the non-dimensional radius. **d** Average non-dimensional concentration of drug versus time for bolus injection

calculated. Tumor with necrotic core of 50% is considered as baseline model ($R_n = 0.5$). Figure 3a–c illustrate the spatial distribution of non-dimensional IFP, IFV, and non-dimensional concentration over the non-dimensional radius, respectively. The maximum value of IFP is obtained at the tumor center and is significantly maintained at the major part of the tumor radius until it begins to decrease sharply at approximately one-tenth of the outer edge of the tumor. Here, the maximum value of IFP is 1,530 Pa. This value is matched to the values reported by Boucher et al. [16], Soltani et al. [17], and experimental results of Huber et al. [18]. The interstitial fluid value is straightly proportional to the local pressure gradient, based on the Darcy equation. Therefore, the IFV magnitude is negligible for $0 < r < 0.9r_t$ (Fig. 3b). However, the IFV sharply increased as a result of steep pressure gradient in the vicinity of the tumor boundary. Thus, the maximum amount of IFV ($0.03 \mu\text{m/s}$) occurs on the tumor surface ($r = r_t$) where radially extends outwards and opposes the drug penetration into the tumor during the chemotherapy. Therefore, drug agents are expected to be significantly hindered at the tumor periphery. After that, IFV decreased to a zero value in normal tissue because of the lack of pressure gradient in this.

Two bumps are observed in non-dimensional concentration distribution at inner boundary 1 and 2. Normal tissue has uniform distribution except near the inner boundary 1 and outer boundary. Non-dimensional concentration value in major part of necrotic core is negligible ($0 < r < 0.45 r_i$) and then it increases. Generally, non-dimensional concentration in the inner boundary 1 has its maximum value.

Figure 3d shows the average non-dimensional concentration of drug versus time for bolus injection. Concentration of drug agents increases rapidly to maximum values about 7 h post injection and decreases thereupon until it approaches zero after 72 h.

The effect of necrotic core size on drug delivery into the tumor has been investigated for three tumors with different sizes, 1, 5, and 10 mm in radius. Figure 4a–c shows the effect of necrosis size on IFP. As can be seen from these figures, for each tumor size, IFP in tumor decreases with increasing R_n . This decrease in IFP becomes more significant at larger R_n values. Another point to be drawn when comparing the results is that IFP in larger tumors has less dependency on necrotic core size. For example, in the 10 mm tumor, no significant change in IFP was observed for R_n values less than 0.9, whereas in the 1 mm tumor at $R_n = 0.9$, IFP changed drastically and reached a value of 0.6 of maximum IFP in tumor.

Figure 4d–f shows the effect of necrotic core size on average non-dimensional concentration distribution of drug for different size of tumors. It is expected that with the increase in R_n which leads to a decrease in the IFP in tumor drug delivery to the tumor will improve. However, as shown in the Figs. 8d–f, with increasing R_n , average non-dimensional concentration of drug in the tumor decreased for all tumors. In fact, tumors with larger values of R_n , have a smaller vascularized area and less effective vessels. For this reason, these tumors potentially have lower chance for drug supply than those with bigger vascularized area. Although a reduction in IFP is considered as a positive factor in improving drug delivery, it should be noted that there is a better drug supply for tumors with smaller necrotic zone which improves drug delivery for this tumors. In summary, tumors with a smaller necrotic core receive a higher level of drug concentration in chemotherapy, although they have higher IFP values.

4 Conclusions

The specific pathophysiology of the tumor results in an elevation in IFP in the tumor, which leads to an outward convection flow that hinders both the convective and diffusive mechanisms of drug delivery. In the present study, a computational model is employed in a solid tumor to calculate the IFV, IFP, and solute transport. The computational approach considered the mechanisms of convection and diffusion of drug in interstitium and drug extravasation from microvessels or to lymphatic vessels. The transient distribution of drug concentration is calculated based on the IFV and IFP distribution. The results show that the semi-necrotic region of the tumor has the maximum value of concentration while the concentration in the necrotic core steeply approaches zero. Increasing R_n values lead to a decrease in IFP in the tumor, which

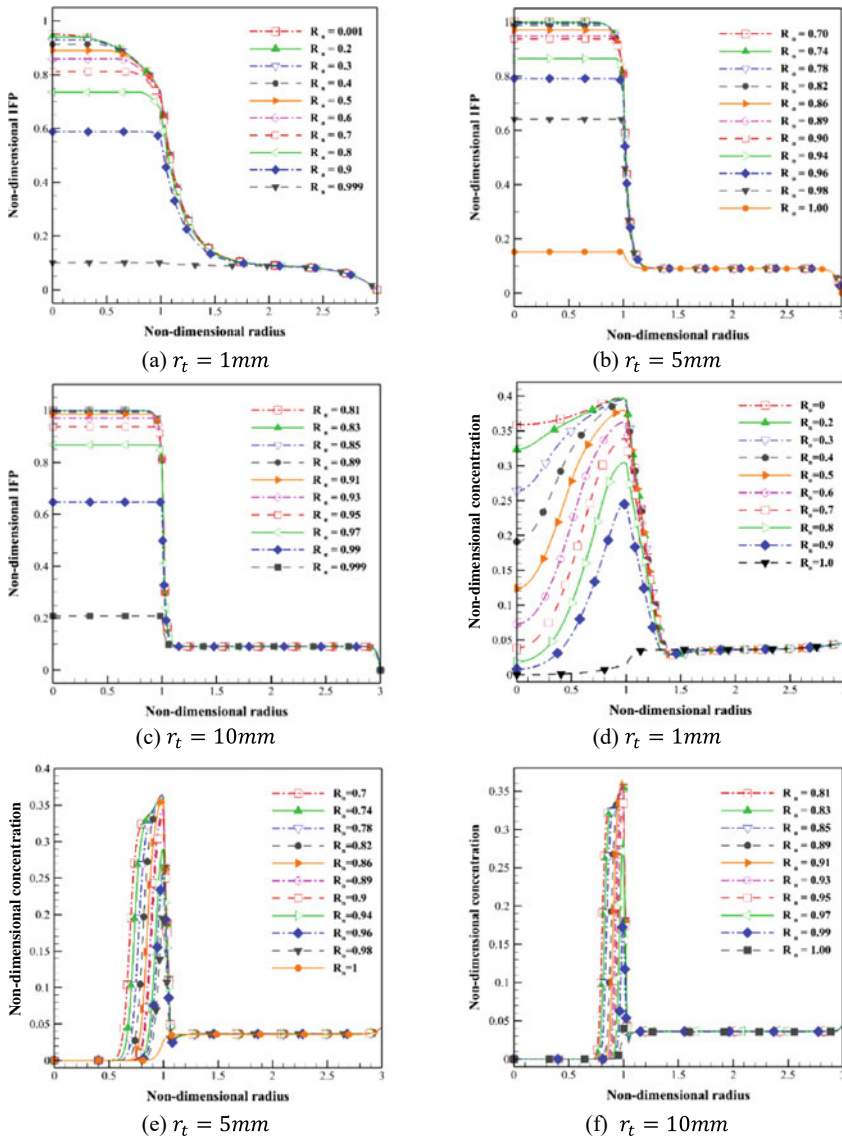


Fig. 4 The effect of necrotic core size on drug delivery. **a–c** Non-dimensional IFP distribution, and **d–f** average non-dimensional concentration distribution of drug for three different sizes of tumors

is more significant at larger R_n values. IFP in smaller tumors are more sensitive to R_n , and even for small values of R_n , a notable decrease in IFP is observed. Although a decrease in IFP makes the drug delivery to the tumor easier, the delivery of drugs to solid tumors with larger necrotic core is decreased due to the reduction of drug supply through the capillary network by the expansion of the necrotic area. Results show that

tumors with a smaller necrotic core receive a higher level of drug concentration, while they have higher IFP values.

References

1. Kashkooli, F.M., et al.: Drug delivery to solid tumors with heterogeneous microvascular networks: novel insights from image-based numerical modeling. *Eur. J. Pharmaceut. Sci.* 105399 (2020)
2. Rezaeian, M., Sedaghatkish, A., Soltani, M.: Numerical modeling of high-intensity focused ultrasound-mediated intraperitoneal delivery of thermosensitive liposomal doxorubicin for cancer chemotherapy. *Drug Delivery* 26(1), 898–917 (2019)
3. Kashkooli, F.M., et al.: Image-based spatio-temporal model of drug delivery in a heterogeneous vasculature of a solid tumor—Computational approach. *Microvasc. Res.* 123, 111–124 (2019)
4. Kashkooli, F.M., Soltani, M., Souri, M.: Controlled anti-cancer drug release through advanced nano-drug delivery systems: static and dynamic targeting strategies. *J. Cont. Rel.* (2020)
5. Baxter, L.T., Jain, R.K.: Transport of fluid and macromolecules in tumors. II. Role of heterogeneous perfusion and lymphatics. *Microvasc. Res.* 40(2), 246–263 (1990)
6. Soltani, M., Chen, P.: Numerical modeling of fluid flow in solid tumors. *PLoS ONE* 6(6), e20344 (2011)
7. Soltani, M.: Numerical modeling of drug delivery to solid tumor microvasculature (2013)
8. Soltani, M., Chen, P.: Numerical modeling of interstitial fluid flow coupled with blood flow through a remodeled solid tumor microvascular network. *PLoS ONE* 8(6), e67025 (2013)
9. Sefidgar, M., et al.: Effect of tumor shape, size, and tissue transport properties on drug delivery to solid tumors. *J. Biol. Eng.* 8(1), 12 (2014)
10. Soltani, M., Chen, P.: Effect of tumor shape and size on drug delivery to solid tumors. *J. Biol. Eng.* 6(1), 4 (2012)
11. Sedaghatkish, A., et al.: Acoustic streaming and thermosensitive liposomes for drug delivery into hepatocellular carcinoma tumor adjacent to major hepatic veins; an acoustics–thermal–fluid-mass transport coupling model. *Int. J. Therm. Sci.* 158, 106540 (2020)
12. Soltani, M., et al.: Effects of magnetic nanoparticle diffusion on microwave ablation treatment: a numerical approach. *J. Magnet. Magnet. Mater.* 167196 (2020)
13. Kashkooli, F.M., et al.: Effect of vascular normalization on drug delivery to different stages of tumor progression: in-silico analysis. *J. Drug Deliv. Sci. Technol.* 101989 (2020)
14. Steuperaert, M., et al.: Mathematical modeling of intraperitoneal drug delivery: simulation of drug distribution in a single tumor nodule. *Drug Delivery* 24(1), 491–501 (2017)
15. Curry, F.-R.E.: Mechanics and thermodynamics of transcapillary exchange. *Handbook of Physiology*, pp. 309–374 (1984)
16. Boucher, Y., Baxter, L.T., Jain, R.K.: Interstitial pressure gradients in tissue-isolated and subcutaneous tumors: implications for therapy. *Can. Res.* 50(15), 4478–4484 (1990)
17. Soltani, M., et al.: Spatiotemporal distribution modeling of PET tracer uptake in solid tumors. *Ann. Nucl. Med.* 31(2), 109–124 (2017)
18. Huber, P.E., et al.: Trimodal cancer treatment: beneficial effects of combined antiangiogenesis, radiation, and chemotherapy. *Can. Res.* 65(9), 3643–3655 (2005)

Evaluating a Logistic K-mer Based Model for Classifying CO1 Sequences of *C. Clupeaformis*



D. St Jean, Herb Kunze, and D. Gillis

Abstract Lake whitefish (*Coregonus clupeaformis*) are a primary support for subsistence and commercial fishing in Canada. In the 20th century, lake whitefish populations experienced a dramatic decline as a result of overfishing, environmental degradation, and predation. With proper environmental management and fishery management, populations have recovered, however certain local populations are still at risk. To properly manage these fisheries for sustainable yield, it is important that the genetic diversity of the population is maintained to ensure evolutionary potential of the species. The expensive technique of physically sampling populations is being replaced by sampling environmental DNA (eDNA) from the physical environment. However, the process of labelling eDNA sequences on a species level is still being developed. We found that techniques and theories from the well-established field of natural language processing, combined with machine learning algorithms, were extremely well-suited to labelling eDNA sequences. We built a logistic k-mer based learning model, inspired from natural language processing, which had high levels of classification accuracy. We anticipate this model is a starting point for more sophisticated machine learning models, and we have demonstrated how processes from the field of natural language processing are analogous to our eDNA process.

Keywords Machine learning · Logistic k-mer model · Natural language processing · eDNA sequences

D. S. Jean (✉) · H. Kunze · D. Gillis
University of Guelph, 50 Stone Rd. E, Guelph, Canada
e-mail: dstjean@uoguelph.ca

H. Kunze
e-mail: hkunze@uoguelph.ca

D. Gillis
e-mail: dgillis@uoguelph.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_56

611

1 Introduction

Recent advancements in DNA based sequencing techniques have led to the ability to rapidly and relatively inexpensively sequence genetic genomes [1]. DNA identification techniques based on genome sequencing can be used to authenticate food that has lost phenotypic characteristics, such as in fillets or other uncooked seafood products. Deliberate or unintentional, seafood fraud is a current issue in Canada, with a recent study showing that over 44% of samples collected from retailers and restaurants were mislabelled [2]. Beyond economic concerns, seafood fraud creates food safety and health risks, threatens oceans, cheats honest fishers and vendors, and creates a market for illegally caught fish, which can mask global human rights abuses [2–4]. As DNA sequencing has advanced, a larger number of samples are able to be processed, leading to a large output of sequenced samples. The Canadian Food Inspection Agency (CFIA) is currently committed to improving how these sequence samples are classified according to genus and species level classifiers [5].

The *Coregonus* genus, belonging to the salmon family, contains at least 68 described species [6]. These species are typically genetically similar, which leads to issues when attempting to classify them into species groups. Within the *Coregonus* genus, the species *C. clupeaformis*, holds value commercially, socially, and ecologically to Indigenous and non-Indigenous peoples [7].

Major techniques used to classify species include phenotype-based classifying, standard genetic techniques, and computer science based approaches [8]. Since phenotypes are lost when seafood is processed, and standard genetic techniques are often time consuming, the development of computer techniques to classify genetic samples is highly valuable [9]. The use of machine learning techniques as an alternative to traditional genetic classification techniques has recently been proposed [10, 11].

This paper will explore the use of a natural language processing (NLP) based machine learning model to classify data from the Barcode of Life (BOL) dataset, and determine the model’s effectiveness in classifying samples.

In Sect. 2, we will explore the inspiration for the model, along with details of the important features of this model. Following that, we will outline the methods and implementation details in Sect. 3, and provide results and discussion in Sects. 4 and 5 respectively.

2 Model Inspiration and Background Information

2.1 Inspiration from Natural Language Processing

DNA sequences are a meaningful genetic language [12] that convey important information about how life functions. This genetic language is analogous to natural language, which humans developed naturally through use [12]. The field of natural language processing (NLP) is concerned with interactions between computer and

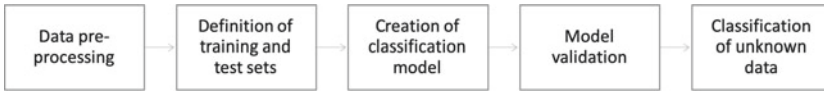


Fig. 1 Typical process for a document classification problem in NLP

human (natural) languages. One classic NLP problem is of document classification; a specific example would be classifying new library books into genres.

In a traditional document classification problem, as shown in Fig. 1, the set of documents to be classified is first pre-processed. Documents are split up into words, sometimes referred to as “tokens”. Unimportant words are then filtered out, this normally is the “stop” words (and, the, or, etc) and extremely rare words that occur in very few documents. In the pre-processing step, other processes such as stemming (reducing words to their base meaning) and sentence boundary detection also take place. In the second step, training and test sets are defined, both of which often come from a pre-existing corpus: a large and structured set of texts that are pre-labelled. Once the documents have been pre-processed and divided, it can be used to train a classification algorithm, of which there are a large variety. Model validation is used to ensure the results from the chosen algorithm perform as expected, and this step may also involve quantifying the robustness and quality of the model. Finally, this model can be used to classify new documents.

What is extremely exciting is that this process of classifying documents from the well-established field of NLP is extremely analogous to the classification of genetic sequences. In this work, we build a prototype model, using the identical process from NLP. Each genetic sequence (“document”) contains 650 base pairs (“letters in a word”). Genetic sequences can have many similar patterns contained within them, similar to how books in related genres contain similar words. In this model, the genetic sequence is split up into words using a technique known as a k-mers-based approach, and word frequency per document (species instance) is then used to train a logistic model. By taking an NLP-based approach to understanding genetic language, many of the well-established techniques from the field of NLP can be applied to genetic sequences.

2.2 *K-mers Based Model*

An important distinction between natural language and genetic language is that a genetic language does not contain a standard vocabulary: each genetic sequence is completely unique. This means that the processing relies heavily on pattern recognition and pattern matching. In computational genetics, k-mers refers to all possible subsequences (of length k) from a genetic sequence [13]. For example, the genetic sequence TCGATT has four unique 3-mers: TCG, CGA, GAT and ATT.

This implementation of the model looked exclusively at 30-mers, as very low level mers are extremely common within all sequences [14], and higher count *mers* are too rare to occur in multiple sequences, similar to how word frequency varies according to Zipfs law [15].

The advantage to using a k-mer based segmentation of genetic sequences compared to traditional comparisons of genetic sequences is that it eliminates the need for alignment. Traditionally, to compare two genetic sequences, the sequences would be aligned at a spot that was assumed to be similar (for example: an important nucleotide) and then the remaining base pairs in the genetic sequence would be directly compared for similarity. By using a k-mer approach, the alignment component is eliminated, and the problem becomes one of pattern recognition. This can be considered equivalent to classifying similar documents that contain certain words or certain substrings of words.

One disadvantage to using a k-mers based approach is that genetic sequences are often very similar, which may lead to a high overlap in subsequences; this is comparable to the high frequency of common words in natural language. This can be overcome by filtering out subsequences that are very common, or by only considering longer k-mers.

3 Methods and Implementation Details

3.1 Logistic Regression

Once all of the genetic sequences were segmented into words, a logistic regression model was trained. Logistic regression was chosen for this problem because the dependent variable was binary: the instances either belonged to *C. clupearformis*, or they did not. The independent variables are each unique 30-mers, and each species instance either tests positive (1) or negative (0) for that specific k-mer. The total number of unique 30-mers from all sequences was 235,279 with each sequence containing an average of approximately 626 unique 30-mers.

Logistic regression takes the unique 30-mers and weights them based on how well they predict that a genetic sequence belongs to *C. clupearformis*. The logistic regression models the probability that the observed data belong to *C. clupearformis*, or not, based on which of the independent variables (unique 30-mers) are present/absent in each genetic sequence. The logistic regression model implemented came from scikit-learn (sklearn) a free library [19] for Python [21].

To determine y , the probability of success (that is, that an observation belongs to *C. clupearformis*), we use the following equation,

$$y = \frac{e^{\beta_0 + \beta x}}{1 + e^{\beta_0 + \beta x}} \quad (1)$$

Here, \mathbf{x} is a vector of 1s and 0s indicating which of the unique 30-mers belong to each observation, and β their corresponding weights.

In this model, each unique 30-mer has an associated β coefficient that is determined from the training data, using maximum-likelihood estimation. The output of the model is the probability that an input sequence belongs to the class *C. clupearformis*, given whether or not the input sequence contains or does not contain each unique 30-mer. More formally, this can be written as $P(\text{sequence} = C. \text{clupearformis} | \mathbf{x})$.

An advantage to using a logistic regression model is that the probability that an instance belongs to a class is outputted, which provides information about how confident the model is that a certain instance belongs to a class. However, since the coefficients that determine the weight of each feature are estimate using maximum-likelihood, an artifact of this method is that no one feature will be weighted to 100% separate the two classes: this means that even if a certain feature has 100% prediction power, the model will not fully utilize this.

3.2 K-Fold Cross-Validation

An important design consideration was taking into account the relatively small data set: the BOL project contains only 328 *Coregonus* records from Canada. There was also the fact that 68.6% of these records belonged to *C. Clupearformis*, while the remaining records came from nine other *Coregonus* species. This means that the classes (species clusters) within the data set are not proportionate, which may introduce bias when training the model.

To combat the size of the data set as well as the stratification of the data set, k-fold cross validation was implemented. Cross-validation is a re-sampling procedure that is used to estimate the proficiency of a machine learning model [16]. It is widely used, because it is fairly easy to understand, and it generally results in a less biased or less optimistic estimate of the model skill [17]. In a traditional train/test split, a set amount of the data is sampled for training, and the rest is carried over to test the trained model. With k-fold cross validation, different train/test splits are used to train and test k models. K-fold cross validation follows a standard process [16, 17]:

1. Take one section of the data as a test data set
2. Dedicate the remaining data as the training data set
3. Fit a model on the training data and evaluate on the test data set
4. Retain the evaluation score and discard the model
5. Begin step 1 with a different section as a test data set

In this model, a value of $k=10$ folds was chosen. In general, the higher the k value, the less bias there is in the model, but this comes at a cost computationally. An extension to this technique would be to implement stratified k-fold cross validation, to ensure that each fold has the same proportion of instances from *C. clupearformis* group and the other species.

3.3 Implementation Details

This experiment relied heavily on the Python library pandas for data structuring. The FASTA files from the BOL project were loaded into pandas DataFrames. A DataFrame is a dict-like container for series objects, that is a two-dimensional data structure with labeled axes [18]. Python's pandas library is excellent for handling data, and this project made use of built-in functions such as copying, indexing, and splitting to restructure the data into a format suitable for the model.

The 30-mers for each genetic sequence were read into a dictionary, and each of these dictionaries was merged in a new DataFrame. To generate a target DataFrame, a direct string match for *Coregonus clupeaformis* was marked as True (1), otherwise False (0).

The scikit-learn (import as sklearn) was used to train and test the logistic regression model [19]. The scikit-learn library features various classification, regression, and clustering algorithms, as well as supporting k-fold cross validation. The metrics (scoring) of the model was also implemented using the scikit-learn library.

The data set used for this project was obtained from the Barcode of Life project [20]. The Barcode of Life Database (BOLD) is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada [20]. One component of the platform is the registry of BINs (putative species): currently there are over 578,000 BINs [20].

While this model accuracy is important, we must also examine how robust the model performs. We examine metrics other than just accuracy, to have a better sense of how reliable the model is. When evaluating the model, we are concerned with four types of outputs: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The accuracy, precision, and recall metrics combine these output types. Accuracy is often described as how well the model classifies on average. Precision and recall, however, indicate how often our model falsely classifies inputs. For example, a model that classifies all sequences as whitefish would be 100% accurate at classifying whitefish, since all whitefish are captured as whitefish. However, this model would have low precision and recall, since there are likely many fish that do not belong to this class, but they are being incorrectly labelled as whitefish.

$$Accuracy = \frac{TP + TN}{total} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Table 1 Logistic 30-mers model metrics with 10 cross-fold validations

Metric	Value (95% CI)
Accuracy	99.5 (97.9, 100.0)%
Precision	99.2 (96.9, 100.0)%
Recall	100.0 (100.0, 100.0)%
R ² Value	97.7 (90.9, 100.0)%

4 Results

The outcome of this model was the ability to classify *C. clupearformis* at a 99.5 % level of accuracy with a 95% confidence interval of (97.9, 100.0)%. A summary of the model results and metrics is shown in Table 1.

5 Discussion

In this model, the high precision value indicates that the majority of positives generated belong to the class of true positives, which means that our model rarely classifies a non-*C. clupearformis* as *C. clupearformis*. The perfect recall indicates that there were no false negatives: every *C. clupearformis* was correctly identified as *C. clupearformis*. Furthermore, the R-squared value of 97.7 (90.9, 100.0)% means that almost all changes of the dependent variable are completely explained by changes in the independent variables. Specifically, this means that the presence or absence of the 30-mers can very accurately explain whether or not an instance belongs to *C. clupearformis* or not.

While all these metrics depict a model which is accurate, precise, and has high recall, this may be due to over fitting of the model, which means that lower metrics may occur with further testing. Future work will involve using more data (real and simulated) to investigate further the robustness of this model. Next steps will include reducing the dimensionality of the problem, as well as investigating how possible parameter estimation.

6 Conclusion

In this work, the problem of classifying *C. clupearformis* was successfully addressed, using a natural language inspired, k-mers based logistic regression model. One of the main contributions of this project is demonstrating that this technique is not limited

to natural language classification problems and lays the foundation for using more complicated NLP-inspired techniques in the future.

A discussion on the specific techniques implemented was provided, as well as a detailed discussion regarding the metrics of the model that was developed. This was important to ensure that there was adequate confidence in the classification accuracy of the model.

Further extension of this approach may allow for classification of multiple species, through an implementation of a multinomial regression model. This described model will hopefully serve as a prototype for future NLP-based species classification. Future work will include increasing the amount of sequences used for training this model, which will likely give a more complete perspective on the robustness of this model.

Acknowledgements We acknowledge that the University of Guelph resides in the ancestral and treaty lands of the Attawandaron people and the Mississaugas of the Credit, and we recognize and honour our Anishinaabe, Haudenosaunee, and Métis neighbours. This work was supported by the HQP (Highly Qualified Personnel) Scholarship Program. This project was completed with the support of the Hanner Lab at the University of Guelph.

References

1. Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E., Zakharov, E.V.: Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States. *PLoS One*. **9**(4) (2014)
2. Roebuck, K., et al.: Canadians eating in the dark: a report card of international seafood labelling requirements. Ecology Action Centre (2017). Available via CASS Resources. <https://solutionsforseafood.org/cass-resources/canadians-eating-dark-report-card-international-labelling-requirements/>. Cited 2 Nov 2019
3. Spink, J., Moyer, D.C.: Defining the public health threat of food fraud. *J. Food Sci.* **76**(9), 157–163 (2011)
4. Lewis, S.G., Boyle, M.: The expanding role of traceability in seafood: tools and key initiatives. *J. Food Sci.* **82**, 13–21 (2017)
5. Wong, E.H.K., Hanner, R.H.: DNA barcoding detects market substitution in North American seafood. *Food Res. Int.* **41**(8), 828–837 (2008)
6. “Coregonus Clupeaformis.”. *FAO Fisheries I& Aquaculture* (2019). Available via [FAO. www.fao.org/fishery/species/2941](http://www.fao.org/fishery/species/2941). Cited 28 Oct 2019
7. Overdyk, L.M., et al.: Real-time PCR identification of lake whitefish *Coregonus clupeaformis* in the Laurentian Great Lakes. *J. Fish Bol.* **88**(4), 1460–1474 (2016)
8. Cawthorn, D.M., Mariani, S.: Global trade statistics lack granularity to inform traceability and management of diverse and high-value fishes. *Sci. Rep.* **7**(1) (2017)
9. Bernatchez, L., et al.: Harnessing the power of genomics to secure the future of seafood. *Trends Ecol. Evol.* **32**(9), 665–680 (2017)
10. Cordier, T., Forster, D., Dufresne, Y., Martins, C.I., Stoeck, T., Pawlowski, J.: Supervised machine learning outperforms taxonomy based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* **18**(6), 1381–1391 (2018)
11. Gerhard, W.A., Gunsch, C.K.: Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ. Int.* **124**, 312–319 (2019)
12. Ratner, V.A.: The genetic language: grammar, semantics, evolution. *Genetika* **29**(5), 709–719 (1993)

13. Ezziane, Z.: Applications of artificial intelligence in bioinformatics: a review. *Expert Syst. Appl.* **30**(1), 2–10 (2006)
14. Searls, D.B.: The language of genes. *Nat.* **420**(6912), 211 (2002)
15. Zipf, G.K.: *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Boston (1949)
16. A Gentle Introduction to k-Fold Cross-Validation. *Machine Learning Mastery* (2018). Available via: Machine Learning Mastery. machinelearningmastery.com/k-fold-cross-validation/. Cited 1 Nov 2019
17. Cross Validation Explained: Evaluating Estimator Performance. *Towards Data Science* (2018). towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85. Cited 26 Oct 2019
18. McKinney, W.: Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*, pp. 51–56 (2010)
19. Pedregosa, F., Varoquaux, G., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
20. Ratnasingham, S., Hebert, P.D.N.: BOLD: the barcode of life data system (www.barcodinglife.org). *Mol. Ecol. Resour.* **7**, 355–364 (2007)
21. van Rossum, G.: Python tutorial. In: *Technical Report CS-R9526*. Centrum voor Wiskunde en Informatica (1995)

Mathematical Modeling of Coupled Electro-thermal Response of Nerve Tissues Subjected to Radiofrequency Fields



Sundeep Singh and Roderick Melnik

Abstract This study aims at developing a mathematical model taking into account the effects of thermal pain sensation induced during the radiofrequency heating of neural tissues. A three-dimensional heterogeneous computational domain comprising of muscle, bone and target nerve has been considered. Importantly, the main governing equations of the multi-scale and multi-physics model are: (a) a simplified version of Maxwell's equation utilizing a quasi-static approximation for estimating the electric field distribution, (b) the Pennes bioheat transfer equation for estimating the temperature distribution and (c) a modified Hodgkin-Huxley model for prediction of nociceptor electrophysiology. The temperature-controlled radiofrequency has been modeled on the neural tissue by utilizing the protocols applied in actual clinical practices along with taking into account the temperature-dependent electrical conductivity and blood perfusion rate. The effects of different values of preset target temperature on the treatment outcomes of nerve ablation have also been quantified. The findings of the present study would provide *a priori* information to the clinicians that will be beneficial during the treatment planning stage of the therapy.

Keywords Thermal therapies · Pain relief · Nerve ablation · Nociceptor · Thermal pain · Coupled electro-thermal model

1 Introduction

Chronic pain is one of the most common problems affecting millions of Canadians each year and contributes to the significant burden on healthcare resources (\approx \$7.2

S. Singh (✉) · R. Melnik
MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University,
75 University Avenue West, Waterloo, Ontario N2L 3C5, Canada
e-mail: ssingh@wlu.ca

R. Melnik
e-mail: rmelnik@wlu.ca

R. Melnik
BCAM—Basque Center for Applied Mathematics, E-48009 Bilbao, Spain

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_57

billion annually) [1]. The management of chronic pain is largely based on the use of opioids medication, misuse of which could lead to negative effects such as physical dependence and addiction [1, 2]. There has been a continuous quest for exploring treatment options that are cheap and effective for long durations, and could minimize the need for opioids. Radiofrequency ablation is one such treatment options that has been frequently applied in the last decade for the management of different types of chronic pain [3, 4]. The setup of the minimally invasive radiofrequency (RF) procedure comprises of the RF generator, the dispersive ground pad and an electrode with a small portion of the active length. Once the electrode is placed at the target site, alternating current within the kHz range is delivered from the RF generator to the target nerve via the active part of the electrode that is further captured and returned back to the generator by the ground pad, forming a closed electric circuit. As the RF current interacts with the biological tissue, frictional heating is induced due to the agitation of ions present within the tissue electrolytes [2, 4]. By virtue of this, temperatures above 50 °C are achieved close to the active site of the electrode leading to biological changes such as protein and collagen denaturation and ultimately coagulative necrosis. The attainment of high temperature (close to 100 °C) during the RF procedure could further lead to side effects such as tissue charring and vaporization, and is often an indication to stop the RF procedure as it results in a drastic decline in electrical and thermal conductivities of the tissue, thereby acting as a barrier to the efficient conduction of thermal energy [3]. In general, the RF power is delivered to the neural tissue using different modes, viz., continuous and pulsed. In the conventional continuous mode, the RF power is delivered to the target neural tissue in a continuous manner, leading to the destruction of the axons and limiting the transmission of pain signals. While, in the pulsed RF mode, short pulses of RF currents are applied to the target nerve that allows heat to dissipate and restricts the attainment of temperatures above 42 °C during the entire procedure, thereby avoiding any thermal damage [2, 4].

Although the usage of RF procedure for mitigating chronic pain has been increasing tremendously during the past decades, several questions and controversies still prevail regarding the underlying mechanism, efficacy and benefits of RF [1, 2, 4]. Computational modeling can provide a cheap and viable alternative for quantifying the underlying physics and providing *a priori* estimate of the treatment outcomes that could assist the clinicians in optimizing and standardizing the treatment protocols specific to particular target sites of neural tissues. In what follows, the present study focuses on developing more realistic three-dimensional heterogeneous models of continuous RF for treating chronic pain. A temperature-controlled algorithm has been used whereby the maximum temperature during the RF procedure won't be allowed to reach 100 °C to mitigate any chances of occurrence of undesirable phenomena of charring and water vaporization. The effect of preset target temperature on the applied voltage, temperature distribution and ablation volume has been quantified by conducting a coupled thermo-electric analysis. Furthermore, the rise in tissue temperature during RF procedure could also lead to the induction of nociceptive pain other than the target site and mainly close to the skin tissue. Importantly, the transduction of nociceptive pain occurs through the nociceptors that reside at the

ends of the long axons of neurons and mediate the selective passage of specific ions or molecules across cell membranes at noxious temperature levels. The nociceptors are one of the three kinds of peripheral nerves: myelinated afferent $A\delta$ and $A\alpha$ fibers, and unmyelinated C afferent fibers. Thermal pain sensations are mainly mediated by both myelinated $A\delta$ and unmyelinated C fibers [5, 6]. Thus, the effect of such nociception has also been taken into consideration within the computational model of the temperature-controlled RF procedure.

2 Computational Modeling Details

The schematic of a three-dimensional heterogeneous computational domain comprising of muscle, bone and nerve tissue [7] considered in the present numerical study has been presented in Fig. 1. A 22-gauge (5 mm active length) monopolar RF electrode [8] has been inserted parallel to the target nerve, as shown in Fig. 1. The temperature-controlled RF procedure has been performed by utilizing a closed-loop feedback proportional-derivative-integral (PID) controller that continuously modulates the applied voltage at the active length of the electrode on the basis of the difference between the preset target temperature and the predicted temperature at the tip of the RF electrode [9, 10]. Three different values of preset target tip temperatures, viz., 65, 75 and 85 °C have been considered. The dispersive ground electrode has been modeled by applying zero voltage boundary conditions at the outer boundaries of the computational domain. The initial voltage of the computational domain has

Fig. 1 (Color online)
Schematic of a three-dimensional heterogeneous computational domain comprising of nerve, bone and muscle tissue along with monopolar RF electrode

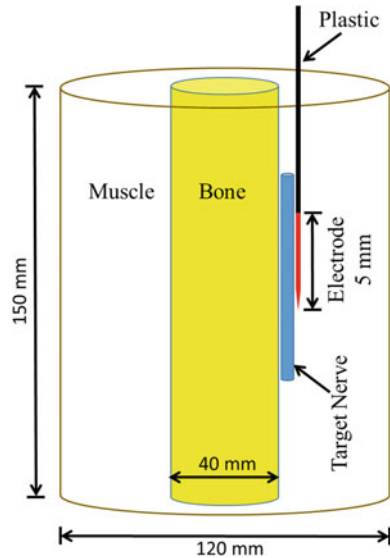


Table 1 Electric, thermal and biophysical properties of different materials considered in this study

Material	σ (S/m)	c (J/kg/K)	k (W/m/K)	ρ (kg/m ³)	ω_b (s ⁻¹)
Muscle	0.446	3421	0.49	1090	6.35×10^{-4}
Bone	0.0222	1313	0.32	1908	4.67×10^{-4}
Nerve	0.111	3613	0.49	1075	3.38×10^{-3}
Plastic	10^{-5}	1045	0.026	70	–
Electrode	7.4×10^6	480	15	8000	–

been considered to be 0 V and the initial temperature has been considered similar to the core body temperature of 37 °C. The material properties considered in the present study are provided in Table 1 [3, 7, 8, 11, 12].

The computational model considered in this study is based on the coupled thermo-electric problem where electromagnetic energy is used to heat the neural tissue during continuous RF procedure for treating chronic pain. Due to the lower frequency range of 450-550 kHz used during the RF procedures, the wavelength of the electromagnetic field is several orders of magnitude larger than the size of the active electrode. Thus, a simplified version of Maxwell's equations, known as the quasi-static approximation, can be used to solve the electromagnetic problem without compromising accuracy [3, 7–12]. The governing equation for the electrical problem is given by:

$$\nabla \cdot (\sigma(T)\nabla V) = 0, \quad (1)$$

where σ (+2 % per °C) is the temperature-dependent electrical conductivity (S/m) [9] and V is the applied voltage (V). Further, the volumetric heat source, Q_p (W/m³), generated by RF currents within the biological tissue is given by:

$$Q_p = \mathbf{J} \cdot \mathbf{E}, \quad (2)$$

where $\mathbf{E} = -\nabla V$ (in V/m) is the electric field and $\mathbf{J} = \sigma(T)\mathbf{E}$ (in A/m²) is the current density.

The governing equation for the thermal problem is the Fourier-conduction-based Pennes bioheat transfer equation and is given by:

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot (k\nabla T) - \rho_b c_b \omega_b (T - T_b) + Q_m + Q_p, \quad (3)$$

where ρ is the density (kg/m³), c is the specific heat capacity (J/kg/K), k is the thermal conductivity (W/m/K), ρ_b is the density of blood (3617 kg/m³), c_b is the specific heat capacity of blood (1050 J/kg/K), ω_b is the blood perfusion rate (1/s) i.e. volume blood per unit mass of tissue per unit time, T_b is the blood temperature (37 °C), T is the unknown temperature of the tissue to be computed from Equation 3, Q_p is the volumetric heat source (W/m³) computed using Eq. 2, Q_m is the metabolic heat generation (W/m³) that has been neglected in the present study [8] due to its

insignificant contribution as compared to Q_p and t is the duration of the temperature-controlled RF procedure (s).

In the present computational study, a temperature-dependent piecewise model of blood perfusion rate has been considered. Accordingly, a constant predefined value of blood perfusion rate has been assumed below the tissue temperature of 50°C and beyond that, the value of blood perfusion rate has been considered to be zero owing to the complete cessation of blood perfusion rate due to the collapse of microvasculature within the tissue [13] and is given by:

$$\omega_b(T) = \left\{ \begin{array}{ll} \omega_{b,0} & \text{for } T < 50^\circ\text{C} \\ 0 & \text{for } T \geq 50^\circ\text{C} \end{array} \right\}, \tag{4}$$

where $\omega_{b,0}$ is the constant blood perfusion rate of tissue provided in Table 1.

The ablation volume (\dot{V}) induced during the temperature-controlled RF procedure for chronic pain relief has been quantified using the isotherm of 50°C (i.e. the volume of tissue within the computational domain having temperature $\geq 50^\circ\text{C}$ after the RF procedure) [13] and is given by:

$$\dot{V} = \iiint_{\Omega} dV \text{ (mm}^3\text{) (where } \Omega \geq 50^\circ\text{C).} \tag{5}$$

A modified Hodgkin-Huxley model has been used for modeling the nociceptor signal transduction induced due to the high temperature attained during the RF procedure close to the skin surface. It is given by [5, 6]:

$$C_{mem} \frac{dV_{mem}}{dt} = I_{st} + I_{Na} + I_K + I_{K2} + I_L, \tag{6}$$

where V_{mem} is the membrane potential (mV) that is positive for depolarized membrane and negative for hyperpolarized membrane, t is the neuronal discharge time (ms), C_{mem} is the membrane capacitance per unit area ($\mu\text{A}/\text{cm}^2$). I_{Na} , I_K and I_L are the sodium, potassium and transmembrane leakage currents (all in $\mu\text{A}/\text{cm}^2$), respectively, while I_{K2} is the fast transient potassium current. They are computed as follows:

$$I_{Na} = g_{Na} m^3 h (V_{Na} - V_{mem}); \quad I_K = g_K n^4 (V_K - V_{mem}), \tag{7}$$

$$I_L = g_L (V_L - V_{mem}); \quad I_{K2} = g_A A^3 B (V_{K2} - V_{mem}), \tag{8}$$

where V_{Na} , V_K , V_L and V_{K2} are the corresponding reversal potentials (in mV) for the sodium, potassium, leakage and fast transient sodium currents, respectively, g_{Na} , g_K , g_L and g_{K2} (in mS/cm^2) are the maximum ionic conductances per unit area through the sodium, potassium, leakage and fast transient sodium current components, respectively; m , n and h are the gating variables, and A and B are factors having the same functional significance as factors m and h . $I_{st} = I_{mechanical} + I_{heat} + I_{chemical}$ is the

total stimulation induced current (in $\mu\text{A}/\text{cm}^2$) that can be computed as the sum of the currents generated due to the opening of mechanically-, thermally- and chemically-gated ion channels, respectively. Since in this study only the thermal stimulation was applied on the axons, thus only thermally-gated ion channels were considered for the generation of stimulation current [6] and is given by:

$$I_{st} = I_{heat} = \left(\left[C_{h1} \exp\left(\frac{(T - T_{thr})/T_{thr}}{C_{h2}}\right) + C_{h3} \right] + I_{shift} \right) \cdot H(T - T_{thr}), \quad (9)$$

where T is the temperature experienced by nociceptors, T_{thr} is the thermal pain threshold temperature, C_{h1} , C_{h2} and C_{h3} are constants and I_{shift} is the shift current that ensures that the action potential is generated when $T \geq T_{thr}$ while none is generated if $T < T_{thr}$. H is the Heaviside function accounting for the threshold process. The gating variables: m , n and h can be computed from the following equations [6]:

$$\frac{dx}{dt} = \alpha_x (1 - x) - \beta_x x, \quad (10)$$

$$\alpha_n = -0.01 (V_{mem} + 50) / (\exp[-(V_{mem} + 50)/10] - 1), \quad (11)$$

$$\alpha_m = -0.1 (V_{mem} + 35) / (\exp[-(V_{mem} + 35)/10] - 1), \quad (12)$$

$$\beta_n = 0.125 \exp[-(V_{mem} + 60)/80]; \quad \beta_m = 4 \exp[-(V_{mem} + 60)/18], \quad (13)$$

$$\alpha_h = 0.07 \exp[-(V_{mem} + 60)/20]; \quad \beta_h = 1 / (\exp[-(V_{mem} + 30)/10] + 1), \quad (14)$$

where x is one of the three gating variables (m , n or h), α_x and β_x are the rate constants (s^{-1}) determined from the voltage clamp experiments [6]. Further, the factors A and B are determined from the following sets of equations [6]:

$$\tau_A \frac{dA}{dt} + A = A_\infty, \quad (15)$$

$$A_\infty = \left(0.0761 \frac{\exp[(V_{mem} + 94.22)/31.84]}{1 + \exp[(V_{mem} + 1.17)/28.93]} \right)^{1/3}, \quad (16)$$

$$\tau_A = A_{fac} \left(0.3632 + \frac{1.158}{1 + \exp[(V_{mem} + 55.96)/20.12]} \right), \quad (17)$$

$$\tau_B \frac{dB}{dt} + B = B_\infty, \quad (18)$$

$$B_\infty = \left(\frac{1}{1 + \exp[(V_{mem} + 53.3)/14.54]} \right)^4, \quad (19)$$

$$\tau_B = B_{fac} \left(1.24 + \frac{2.678}{1 + \exp[(V_{mem} + 50)/16.03]} \right). \quad (20)$$

Motivated by [6], the values of different parameters used in the nociception model are: $C_{mem} = 2.8 \mu\text{F}/\text{cm}^2$, $g_{K2} = 47.7 \text{ mS}/\text{cm}^2$, $g_{Na} = 120 \text{ mS}/\text{cm}^2$, $g_K = 36 \text{ mS}/\text{cm}^2$, $g_L = 0.3 \text{ mS}/\text{cm}^2$, $A_{fac} = B_{fac} = 7.0$, $V_{Na} = 57.19 \text{ mV}$, $V_K = -78.78 \text{ mV}$, $V_L = -63.79 \text{ mV}$, $C_{h1} = C_{h2} = 2$, $C_{h3} = -1 \mu\text{A}/\text{cm}^2$, $T_{thr} = 43^\circ\text{C}$ and $V_{rest} = -70 \text{ mV}$.

The coupled thermo-electric models of temperature-controlled RF procedure for treating chronic pain have been solved by the Finite Element Method (FEM) using COMSOL Multiphysics 5.2 software [14] utilizing an adaptive time-stepping

scheme. The computational domain has been discretized using heterogeneous tetrahedral mesh, comprising of 174486 elements and 476384 degrees of freedom, constructed with COMSOL's built-in mesh generator. A further refinement in the area surrounding the active tip of the electrode has been applied, where the highest electrical and thermal gradients are expected. A mesh convergence analysis has been carried out to determine the optimal number of mesh elements that would result in a mesh-independent solution. The temperature distribution computed from the coupled thermo-electric model was fed in the MATLAB code of the modified Hodgkin Huxley model for predicting the nociceptor response to these predicted temperatures. All simulations were run on a Dell T7400 workstation with Quad-core 2.0 GHz Intel® Xeon® processors.

3 Results and Discussion

The effect of different values of preset target temperature, viz., 65, 75 and 85 °C, on the tip temperature and the applied voltage during the temperature-controlled RF procedure has been presented in Fig. 2. As it is evident from Fig. 2, initially, the applied voltage value increases monotonically till the preset target temperature

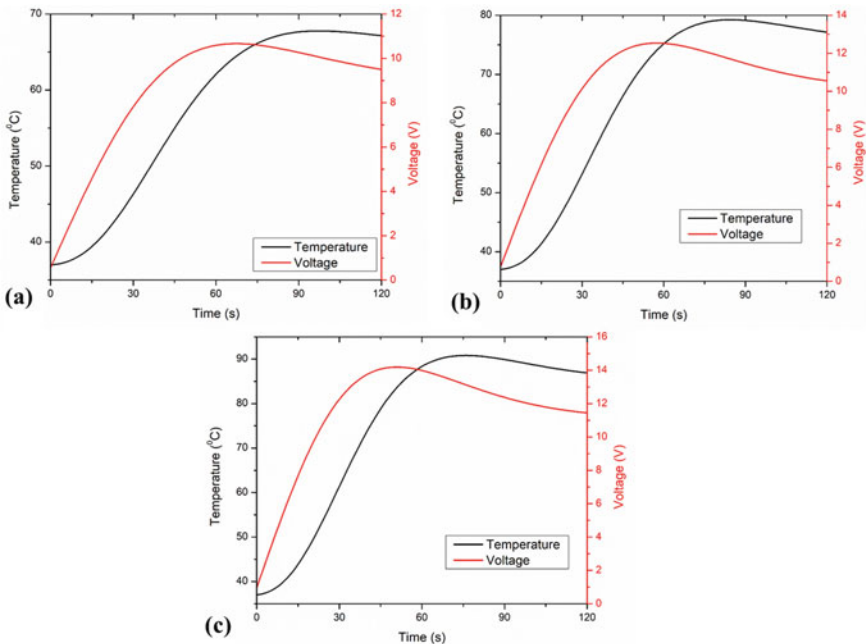


Fig. 2 (Color online) Variation of the applied voltage (red) and tip temperature (black) with respect to time for different values of preset target temperature: **a** 65 °C, **b** 75 °C and **c** 85 °C

has been attained and afterward it declines to maintain the preset value of target temperature. This is true for all values of preset target temperature considered in this numerical study, although the rise in the values of preset target temperature results in the corresponding rise in the applied voltage profile. This can be attributed to the requirement of more energy for attaining the higher target temperature value in comparison to attaining the lower value of target temperature. The maximum values of applied voltages for the preset target temperature of 65, 75 and 85 °C have been found to be 10.66 V, 12.54 V and 14.19 V, respectively. Furthermore, the time required to attain the preset temperature of 65, 75 and 85 °C has been found to be 66 s, 57 s and 51 s, respectively. The variation of target tip temperature follows a similar trends to that of applied voltage, whereby the temperature rises from the core body temperature of 37 °C, i.e. the initial temperature within the computational domain, to the preset value of target tip temperature with an overshoot of $\pm 5\%$ which is common in clinical procedures.

Figure 3 presents the comparative analysis of the total ablation volume (within the entire computational domain) and nerve ablation volume corresponding to the isotherm of 50 °C for different values of the target temperature. The total ablation volume after 120 s of the temperature-controlled RF procedure has been found to be 110.86, 222.79 and 357.50 mm³ for the target temperature values of 65 °C, 75 °C and 85 °C, respectively. Similarly, the damage that occurred to the target nerve tissue alone has been found to be 24.16, 48.91 and 75.48 mm³ for the above target temperature values, respectively, after 120 s of the temperature-controlled RF procedure. Not only this, the variations have also been found in the time at which the initiation of damage occurs for different values of target temperature that basically decreases with the increase of the target tip temperature. Thus, the efficacy of temperature-controlled RF procedure for treating chronic neural pain is significantly dependent on the preset target temperature.

Figure 4 presents the variation of temperature distribution for different preset values of target tip temperature within the computational domain after 120 s of temperature-controlled RF procedure. As depicted in Fig. 4, the attainment of critical temperatures above 50 °C is not only confined to the target nerve, but also to a considerable portion of the healthy muscular tissue on the opposite side of target nerve and bone. The exposure of the muscular tissue just beneath the skin tissue to such higher temperatures could lead to the transduction of nociceptive pain signals through the nociceptors of peripheral nerves (viz., myelinated afferent A δ and A α fibers; and unmyelinated C afferent fibers) residing at the ends of the long axons of neurons. Thus, the present study also models the effects of such high temperature attained on the healthy muscular tissue during the temperature-controlled RF procedure for chronic pain relief. Figure 5 presents the membrane potential and frequency responses under different values of stimulus temperature, viz., 43, 45 and 50 °C. It can be clearly observed from Fig. 5 that the frequency of action potential spikes increases as the nociceptor temperature increases from 43 and 50 °C. The pain level induced due to nociceptor temperature is decided by this signal frequency (i.e. action potential spikes), and thus, the thermal pain level increases as the temperature increases. Such *a priori* estimates about the transduction of nociceptive pain induced due to thermal

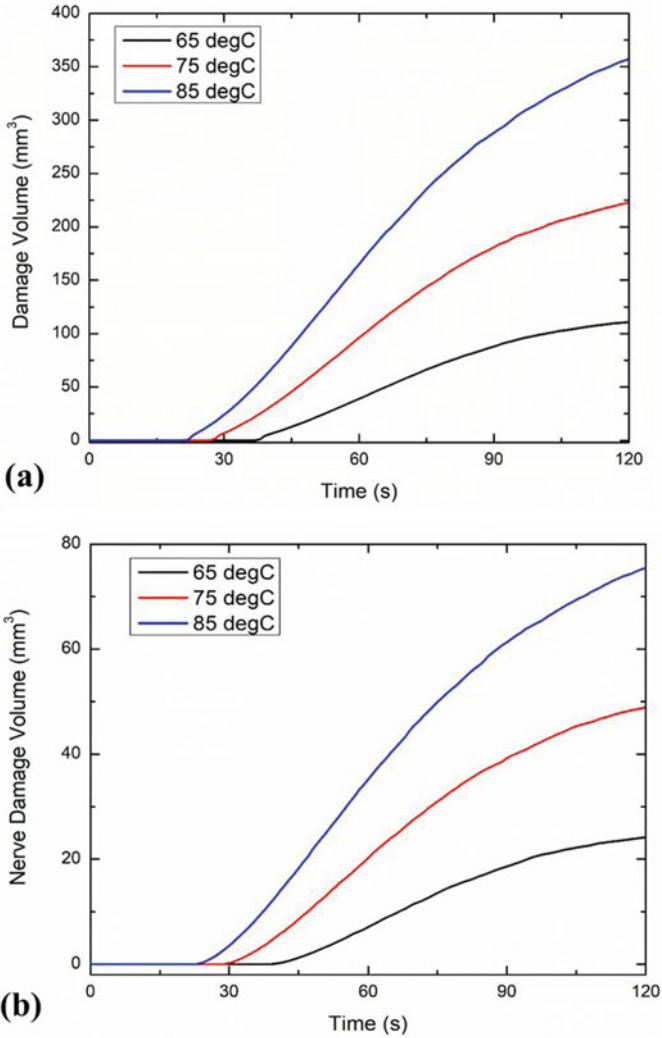


Fig. 3 (Color online) Variation of **a** total damage volume, and **b** nerve damage volume, with respect to time for different values of preset target temperature during the temperature-controlled RF procedure for chronic pain relief (on the insert, black: 65 °C, red: 75 °C, blue: 85 °C)

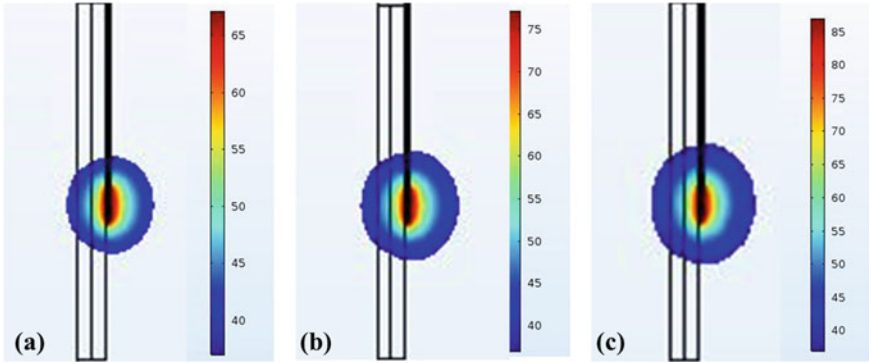


Fig. 4 (Color online) Temperature distribution (in °C) obtained after 120 s of temperature-controlled RF procedure for different values of preset target temperature: **a** 65 °C, **b** 75 °C and **c** 85 °C

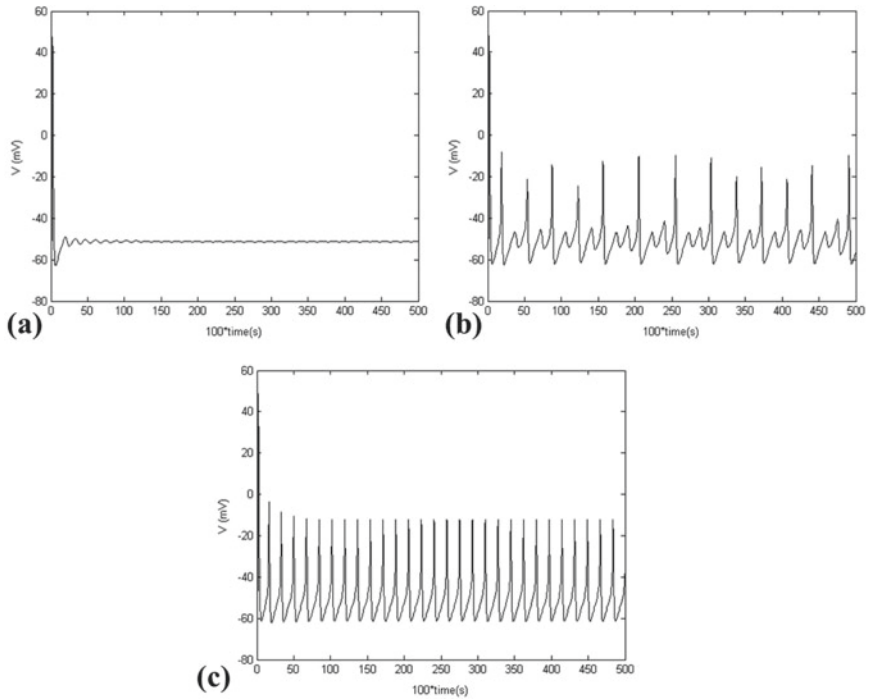


Fig. 5 Predictions of the thermo-neural response of nociceptors at different values of stimulus temperature, viz., **a** 43 °C, **b** 45 °C and **c** 50 °C

stimuli would assist the medical practitioners in designing the anesthesia-free thermal ablation procedures for chronic pain relief.

Future studies will be focused on developing fully coupled thermo-electro-neuronal models by taking into consideration the effect of temperature on the membrane conductance in the Hodgkin-Huxley model and non-Fourier heat transfer, along with the incorporation of actual nerve damage models accounting for decrement of the pain signals when exposed to RF procedures [15, 16]. This will enhance the accuracy of the predictive outcomes from the current model that can be readily integrated into the hospital workflow during the real-time treatment of chronic pain among actual patients in clinical settings. Moreover, image-based patient-specific models derived from actual patient data could significantly assist in bridging the gap between computational and experimental findings.

4 Conclusion

A coupled thermo-electric analysis has been performed for quantifying the effect of preset target temperature on the treatment outcomes of the temperature-controlled RF procedure for chronic pain relief. The study reported a strong dependence of the preset target temperature on the efficacy of RF procedure during neural ablation. It has been found that the ablation volume increases with an increase in target tip temperature and vice-versa. Further, a thermo-neuronal model has also been developed to quantify the induction of pain sensation in the nociceptors during such procedures. These predictions could be quite useful in designing the anesthesia-free RF procedure for chronic pain relief. We expect that the advancement and extension of the developed model can significantly assist the clinicians in better optimizing and standardizing the thermal dosages required for enabling safe and reliable RF applications for mitigating chronic pain.

Acknowledgements Authors are grateful to the NSERC and the CRC Program for their support. RM is also acknowledging support of the BERC 2018-2021 program and Spanish Ministry of Science, Innovation and Universities through the Agencia Estatal de Investigacion (AEI) BCAM Severo Ochoa excellence accreditation SEV-2017-0718, and the Basque Government fund AI in BCAM EXP. 2019/00432.

References

1. Loh, E., Reid, J.N., Alibrahim, F., Welk, B.: Retrospective cohort study of healthcare utilization and opioid use following radiofrequency ablation for chronic axial spine pain in Ontario, Canada. *Reg. Anesth. Pain. Med.* **44**, 398–405 (2019)
2. Deer, T.R., Pope, J.E., Lamer, T.J., Provenzano, D.: *Deer's treatment of pain: an illustrated guide for practitioners.* Springer, Cham (2019)

3. Singh, S., Melnik, R.: Computational analysis of pulsed radiofrequency ablation in treating chronic pain. In: Rodrigues, J., et al. (eds.) *Computational science—ICCS 2019. Lecture Notes in Computer Science*, pp. 436–450. Springer, Cham (2019)
4. Soloman, M., Mekhail, M.N., Mekhail, N.: Radiofrequency treatment in chronic pain. *Expert Rev. Neurother.* **10**(3), 469–474 (2010)
5. Xu, F., Lin, M., Lu, T.J.: Modeling skin thermal pain sensation: role of non-Fourier thermal behavior in transduction process of nociceptor. *Comput. Biol. Med.* **40**(5), 478–486 (2010)
6. Lin, M., Liu, S., Genin, G., Zhu, Y., Shi, M., Ji, C., Li, A., Lu, T., Xu, F.: Melting away pain: decay of thermal nociceptor transduction during heat-induced irreversible desensitization of ion channels. *ACS Biomater. Sci. Eng.* **3**(11), 3029–3035 (2017)
7. Singh, S., Melnik, R.: Radiofrequency ablation for treating chronic pain of bones: effects of nerve locations. In: Rojas, I., Valenzuela, O., Rojas, F., Ortuño, F. (eds.) *Bioinformatics and biomedical engineering (IWBBIO 2019). Lecture Notes in Computer Science*, pp. 418–429. Springer, Cham (2019)
8. Ewertowska, E., Mercadal, B., Muñoz, V., Ivorra, A., Trujillo, M., Berjano, E.: Effect of applied voltage, duration and repetition frequency of RF pulses for pain relief on temperature spikes and electrical field: a computer modelling study. *Int. J. Hyperthermia* **34**(1), 112–121 (2018)
9. Singh, S., Repaka, R.: Temperature-controlled radiofrequency ablation of different tissues using two-compartment models. *Int. J. Hyperthermia* **33**(2), 122–134 (2017)
10. Singh, S., Repaka, R.: Numerical study to establish relationship between coagulation volume and target tip temperature during temperature-controlled radiofrequency ablation. *Electromagn. Biol. Med.* **37**(1), 13–22 (2018)
11. Pérez, J.J., Pérez-Cajaraville, J.J., Muñoz, V., Berjano, E.: Computer modeling of electrical and thermal performance during bipolar pulsed radiofrequency for pain relief. *Med. Phys.* **41**(7), 071708/11 (2014)
12. Singh, S., Melnik, R.: Domain heterogeneity in radiofrequency therapies for pain relief: A computational study with coupled models. *Bioengineering* **7**(2), 35 (2020)
13. Singh, S., Repaka, R., Al-Jumaily, A.: Sensitivity analysis of critical parameters affecting the efficacy of microwave ablation using Taguchi method. *Int. J. RF Microw. Comput. Aided Eng.* **29**(4), e21581 (2019)
14. COMSOL Multiphysics® v. 5.2. www.comsol.com. COMSOL AB, Stockholm
15. Singh, S., Melnik, R.: Thermal ablation of biological tissues in disease treatment: A review of computational models and future directions. *Electromagn. Biol. Med.* **39**(2), 49–88 (2020)
16. Singh, S., Melnik, R.: Coupled thermo-electro-mechanical models for thermal ablation of biological tissues and heat relaxation time effects. *Phys. Med. Biol.* **64**, 245008 (2019)

Ranking Association Rules from Data Mining for Health Outcomes: A Case Study of Effect of Industrial Airborne Pollutant Mixtures on Birth Outcomes



K. Vu, A. Osornio-Vargas, O. Zaïane, and Y. Yuan

Abstract Association rule mining can be a powerful computational tool for exploring complex interactions between high-dimensional exposures and health outcomes. Given the high-dimensional nature of the data, many complex association rules may be identified. To narrow down on the most important rules for hypothesis-generating and future investigation in the context of health research, we need an objective approach to reduce the ruleset. The ranking is often based on the lift, a widely used measure of association strength in data mining. In this paper, we show why the lift-based ranking is undesirable from a population health perspective. We propose a new approach to select rules obtained from association rule mining. This new approach considers both association strength measured by relative risk and the excessive health burden in the target population. We use a case study of rules mined from industrial airborne pollutant mixtures and birth outcomes, comparing rules selected using our proposed approach to those selected using lift.

Keywords Data mining · Association rule · Lift · Relative risk · Association strength · Health burden

K. Vu (✉) · Y. Yuan
University of Alberta School of Public Health, Edmonton, Canada
e-mail: vu@ualberta.ca

Y. Yuan
e-mail: yyuan@ualberta.ca

A. Osornio-Vargas
University of Alberta Department of Pediatrics, Edmonton, Canada
e-mail: osornio@ualberta.ca

A. Osornio-Vargas · Y. Yuan
University of Alberta Women and Children's Health Research Institute (WCHRI),
Edmonton, Canada

O. Zaïane
University of Alberta Department of Computing Science, Edmonton, Canada
e-mail: zaiane@ualberta.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_58

1 Background

The motivation data set for this work comes from the Data Mining & Neonatal Outcomes (DoMiNO) study, which aims to advance the knowledge on how exposure to low-dose airborne chemical mixtures during pregnancy affects birth outcomes [4]. Exposure to certain chemicals in the air from industrial pollution has been shown to increase the risk of diseases, including adverse birth outcomes (ABO), which is a significant concern for health scientists and medical practitioners [7]. However, the effects of industrial air pollution on ABOs are still inadequately understood with inconsistent findings [7]. Over one hundred distinct industrial chemicals often co-emit into the air, possibly having interactive effects on pregnancy. Classic epidemiological approaches for association analysis are not fully equipped to handle the complexity of high-dimensional and high-order interactions. Spatial association rule mining offers an attractive alternative to explore the complex relationship in this type of investigation [15], e.g. associations between chemical mixtures and birth outcomes.

Association rule mining is a method for discovering interesting relations between variables in high-dimension datasets by analyzing patterns of if-then co-occurrence of antecedent(s) (the “if” part) and a consequent (the “then” part). The if-then associations are called association rules. Association rule mining algorithms use indices of so-called “interestingness” to generate and select association rules from complex high dimensional datasets [13]. In association rule mining, one widely used index is “lift” [17], defined as the ratio of the joint occurrence of an antecedent (A) and consequent (C) to the product of marginal occurrences of A and C, adjusting for the total number of records, i.e. $\frac{P(AC)}{P(A)P(C)}$ [3]. However, ranking association rules by lift have critical drawbacks. A specific concern for health researchers is that lift-based rule ranking may overlook association rules with high prevalence of antecedents (e.g. a mixture of chemicals) that are strongly associated with the consequents (e.g. adverse birth outcomes) [18]. This is undesirable and should be avoided from a population health perspective because rules poorly ranked by lift could be of significant interest in health studies.

In this paper, we propose a new approach for ranking rules obtained from association rule mining in the context of population health. It overcomes the limitations of the lift-based ranking and identifies rules that are most likely consequential for the target population. We illustrate both approaches using the DoMiNO spatial data mining example where the association of mixtures of airborne chemicals and birth outcomes is of interest.

2 Methods

2.1 Definitions and Notations

Key concepts and definitions are introduced below using a two by two contingency table (Table 1). In the remainder of the paper, we will use A for antecedents and C for consequents, which is equivalent to the epidemiology terms of exposure and outcome, respectively.

Table 1 Contingency table based on counts

	Consequent (Yes)	Consequent (No)	Total
Antecedent (Yes)	a	b	a + b
Antecedent (No)	c	d	c + d
Total	a + c	b + d	N = a + b + c + d

In data mining, the association strength is measured with lift, defined as

$$lift_{(C|A)} \stackrel{\text{def}}{=} \frac{P(C|A)}{P(C)} \stackrel{\text{def}}{=} \frac{P(CA)}{P(C)P(A)} = \frac{aN}{(a+c)(a+b)} \quad (1)$$

In epidemiology, the association strength is typically measured with the concept of relative risk (RR) [5]. RR is defined as the ratio of the occurrence of C in subjects who are exposed to A and the occurrence of C in subjects not exposed to A, adjusting for the total number of exposed and non-exposed subjects. It can be expressed as

$$RR \stackrel{\text{def}}{=} \frac{P(C|A)}{P(C|\bar{A})} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \quad (2)$$

RR measures the “effect” of A on a *relative* scale. We have previously shown that RR and lift are closely related in [18], i.e.

$$RR = \frac{(1 - P(A))lift}{1 - P(A)lift} \quad (3)$$

Compared to lift, RR is always numerically further from the null value of 1 in both directions when an association exists between C and A. Figure 1 (reproduced from [18]) visualizes the RR-lift relationship for various level of P(A).

While RR is a relative measure, another important measure in health studies for the “effect” of A uses an *absolute* scale. The *excessive* number of C (e.g. ABOs) attributed to A measures the population health burden of C due to exposure to A [5]. This excessive number is proportional to a concept “attributable excessive burden” (AEB), which is the difference in consequent probability between the presence and

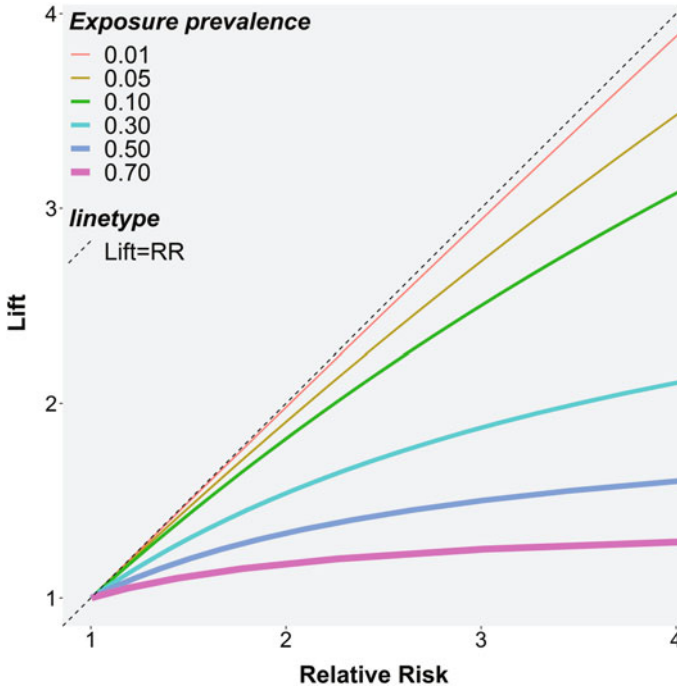


Fig. 1 Relationship between relative risk and lift

absence of antecedent multiplied by antecedent prevalence.

$$AEB \stackrel{\text{def}}{=} P(A)(RR - 1)P(C|\bar{A}) = P(C) - P(C|\bar{A}) = \frac{a + c}{N} - \frac{c}{c + d} \quad (4)$$

High AEB indicates a high burden that could be attributed to A [5]. Thus, in theory, exposures that have higher AEBs should be targeted to reduce the burden of C more effectively.

2.2 Ranking Association Rules

It is beneficial to use a case study to discuss ranking of association rules in a concrete way.

A Case Study

The DoMiNO study used spatial association rule mining to identify mixtures of industrial airborne chemicals associated with ABO, including small for gestational age (SGA), preterm birth (PT), and low birth weight at term (LBWT) [4]. A detailed

description of the data sources and inclusion/exclusion criteria is provided elsewhere [16]. Briefly, population based birth outcomes were obtained from the Alberta Perinatal Health Program (APHP) in the Canadian province of Alberta [2] and included 333,250 singleton live births from 2006 to 2012, of which 29,679 were SGA, 22,733 were PT, and 5,485 were LBWT. The exposure of pregnant women to all registered industrial airborne chemicals ($m = 136$) during the same period was ascertained from industrial emissions data collected by National Pollutant Release Inventory [6] and wind pattern data from 182 stations in Alberta Agriculture's AgroClimatic Information System 2010 [1]. The location of the emission sites, the average emission amount, and the predominant wind (direction and speed) at each site were used to create a dispersion region for each chemical [14]. A pregnant woman was considered exposed to a chemical if her activity area (a 5 km radius from the center of the postal code of her residence) overlapped with the dispersion region of the chemical, as illustrated in Fig. 2 (reproduced from [18]). Each birth and exposure to the chemical(s) form a {chemical(s), birth outcome} transaction for the spatial association rule mining.

The Kingfisher algorithm [8, 9] uses Fisher's exact test to identify significant non-redundant rules for positive associations among all "transactions" using an uncorrected p-value threshold of 0.05. Of the 10,788 rules identified, 2,238, 5,497 and 3,053 were associated with SGA, PT, and LBWT, respectively. The combinations of antecedents included up to 8 chemicals, with lift values ranging from 1.00 to 1.53, corresponding RRs from 1.02 to 1.61, and an extensive range of antecedent prevalence from 0.08 to 98.73%.

As the study aims to explore the relationship between chemical mixtures and birth outcomes, we face the question of how to reduce this large set of more than 10,000 positive rules to a more manageable number.

Problem with lift-based ranking

The conventional data mining approach ranks these rules by lift [17]. However, equation (1) and Fig. 1 demonstrate that the value of lift depends on the prevalence of A ($P(A)$) for a fixed outcome. Consider a hypothetical case where $RR = 4$ (very strong association) and prevalence = 0.95 (extremely high prevalence). In this case, the rule has a corresponding lift of 1.04, which is very close to the null value 1.00, indicating no association. Therefore, ranking by the size of lift would give a low rank to rules that are most consequential to population health and thus the most important to identify.

A new rule selecting method for health outcomes

To select relevant rules using an epidemiological and population health lens, we consider two aspects. First, under the causal framework in health research, the ability to isolate the effect of individual exposures is critical. Relative risk aligns well with the causal framework based on the counterfactual theory [10, 12]. The counterfactual theory states that A causes B if A leads to B, and the difference in the presence of B by the presence versus absence of A is the causal effect. At an individual level, it is impossible to measure this counterfactual effect as we can observe either the presence or absence of A and the corresponding outcome, not both simultaneously. However,

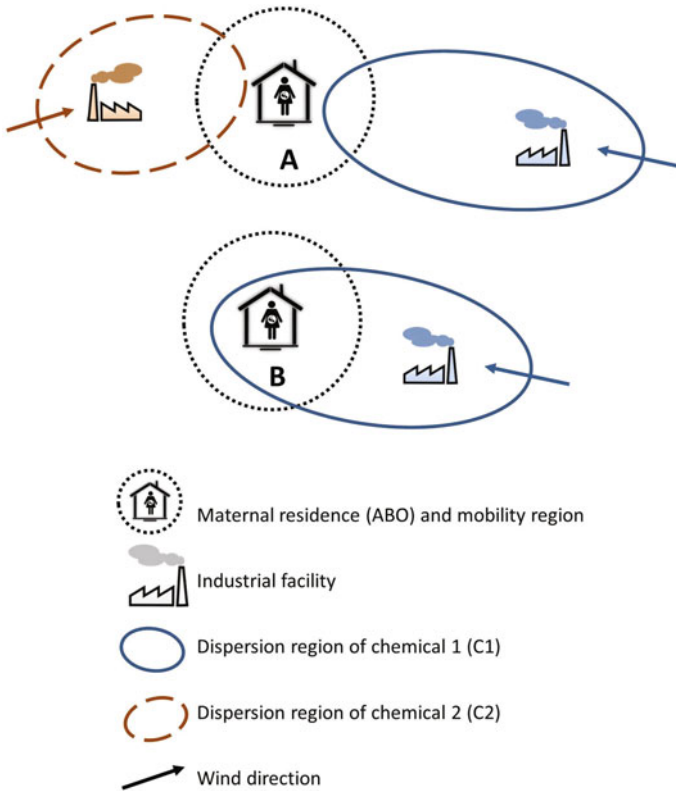


Fig. 2 Illustration of spatial data mining algorithm assigning airborne chemicals exposure to births, which is based on maternal residences, chemical emission sources, and wind information. In this illustration, subject A is exposed to both C1 and C2. Subject B is exposed to C1

the counterfactual effect can be estimated at a population level using comparable groups where the only difference is the presence versus absence of A, such as in a randomized control trial. RR fits this philosophy well, comparing the probability of consequent under the presence versus absence of the antecedent, which differs from lift where the presence of an antecedent is compared to the mixture of its presence and absence [10, 12]. As association strength measured by RR is an essential criterion for causation, a rule with high association strength is more likely a causal relationship, widely accepted in health sciences with work from the well known Bradford Hill [11]. Thus, we suggest converting lift to RR using equation (1) and ranking the rules according to RR and a meaningful association strength threshold, e.g. 1.3, to focus on rules that are more likely causative.

Second, we also consider the absolute “effect” of the antecedents. As the definition for AEB suggests, excessive burden attributable to an antecedent is positively associated with the prevalence of the antecedent $P(A)$. Therefore, to focus on the

rules with a high excessive burden, we suggest imposing a threshold by $P(A)$, e.g. 0.3.

The association rule mining using the Kingfisher algorithm was conducted in Python [8]. The association rule mining results, including lift, support and confidence, were then converted to RR using an R function/SAS macro [18]. which is available on our website, <https://sites.ualberta.ca/~yyuan/software.html>.

3 Results

We used the rules identified from the DoMiNO study to illustrate the disadvantage of ranking by lift and to compare our proposed approach with the lift-based ranking approach. We focused on the 3,053 rules for LBWT, one of the three adverse birth outcomes, as the results from the other two outcomes are similar. In those 3,053 rules, the range of RR went from 1.07 to 1.61, the range of lift from 1.00 to 1.53, and the $P(A)$ varied from 0.4 to 99%.

Table 2 shows the correlation between antecedent prevalence, lift, RR, and AEB of those 3,053 rules. The lift was negatively correlated with AEB, i.e. rules with higher lifts were more likely to have lower AEB ($r = -0.11$). Lift was also negatively correlated with antecedent prevalence ($r = -0.35$).

RR was positively correlated with AEB ($r = 0.43$). Table 2 also shows that RR was weakly correlated with $P(A)$ ($r = 0.19$). We also calculated the lift-RR rank difference. A positive lift-RR rank difference means that a rule was considered more important by RR-based rank than lift-based rank. $P(A)$ was strongly correlated with the lift-RR rank difference ($r = 0.88$), which means that the discrepancy between lift-based rank and RR-based rank became larger with increasing $P(A)$. This illustrates that for rules with a higher antecedent prevalence value (i.e. likely resulting in a higher burden), the lift-based ranking becomes increasingly distant from the RR-based ranking.

As expected, the antecedent prevalence and AEB are highly correlated ($r = 0.96$). This strongly supports the inclusion of $P(A)$ in shortlisting rules for further investigation.

Table 2 Correlations matrix for 3053 rules for low birth weight at term (LBWT)

	P(A)	lift	RR	Lift-RR rank difference ^a	AEB
P(A)	1				
lift	-0.35	1			
RR	0.19	0.85	1		
lift-RR rank difference ^a	0.88	-0.16	0.34	1	
AEB	0.96	-0.11	0.43	0.90	1

^alift-RR rank difference = lift-based rank – RR-based rank

We applied the proposed rule selection method to the rules associated with the LBWT consequent. As an association is considered to exist if $RR \geq 1.2$ [19] and the largest RR for LBWT in the DoMiNO study data was 1.6, we examined the selection method with combinations of RRs (≥ 1.20 , ≥ 1.30 , ≥ 1.35) and $P(A)$ (≥ 0.2 , ≥ 0.3 , ≥ 0.5 , ≥ 0.7). We present the results with $RR \geq 1.3$ and $P(A) \geq 0.3$ in Table 3 as the results from different combinations of RRs and $P(A)$ were consistent.

A total of 358 rules met this criterion. The number of antecedents in the 358 rules ranges from 1 to 7 chemicals. The lift values of the selected rules range from 1.01 to 1.21, leaving out lift values greater than 1.21 (up to 1.53). Table 3 shows that the top rule from this list only ranked 800th according to lift. Lift-RR rank differences ranged from -76 to 2,352, with only eight rules receiving better rank using lift than using RR. As discussed above, one crucial step towards establishing a causal relationship between a specific chemical mixture of interest and LBWT is the strength of the chemical-LBWT association as measured by the relative risk. Exposure to any of the 358 chemical mixtures was associated with a 30 to 35% increase in the risk of LBWT. These relative risk estimates would be the causal effect sizes defined by the counterfactual framework if the causation conditions were met.

On the other hand, it is not straightforward to interpret the lift values of 1.01–1.21 [18]. While we can see associations between exposure to some chemicals and LBWT, due to their corresponding lift values being greater than 1.00, it is not possible to describe the strength of the associations. Notably, some of the lift values associated with the 358 rules were very close to the null value 1.00 (i.e. no association).

These shortlisted rules correspond to AEB values of 0.15 to 0.42% of all 333,250 births, equivalent to 500–1,384 exceeding cases of LBWT, implying that these rules could be “responsible” for 500–1,384 cases in a total of 5,485 LBWT cases. If we rank the 358 rules according to AEB, they occupy the 1st to 1,114th position; Four of these rules are in the top 10, and 10 are in the top 50 AEB-based rules. The eight rules receiving better rank positions with lift than with RR were among the least important rules according to their AEB values. As mentioned, the analyses of different thresholds of RR (1.20, 1.30 and 1.35) and $P(A)$ (0.2, 0.3, 0.5 and 0.7) produced similar results; the rules being removed are those with the highest values of lift. Besides, few or no rules are ranked higher by lift compared to their ranks by RR. This pattern is increasingly more evident as $P(A)$ increases.

The top 100 AEB-based rules for LBWT have high $P(A)$ levels, ranging from 0.57 to 0.99, as expected by the existing high correlation of AEB with $P(A)$. Among these top 100 rules, some correspond to a slightly lower association strength (RR between 1.2–1.3).

4 Discussion

Data mining tools have been increasingly used in health research utilizing linkable massive administrative databases to explore and discover potential associations in a high dimensional setting. It often results in a large number of rules that need to be

Table 3 A summary of rules for LBWT with $RR \geq 1.3$ & $P(A) \geq 0.3$

Rules (n = 358)	Minimum	Maximum
No.chemicals	1	7
P(A)	0.30	0.97
RR	1.30	1.35
RR-based rank	695	1,092
lift	1.01	1.21
lift-based rank	800	3,049
lift-RR rank difference ^a	-76	2,352
AEB (N) ^b	500	1,384
AEB (%)	0.15	0.42
AEB-based rank	1	1,114

^alift-RR rank difference = lift-based rank – RR-based rank

^bNumber of excessive LBWT cases attributable to the effect of antecedent = $N \times AEB(\%)$

reduced to a manageable subset for further investigation. We propose a new selection method designed for rules generated from data mining for health outcomes.

The conventional lift-based rule ranking approach may be problematic when studying health outcomes. A foundation of health research is the causal framework. A measure of association strength, such as lift, is highly undesirable when it negatively correlates with P(A) and the attributable excessive burden.

Our proposed rule selection method rests on RR and the prevalence of antecedents. RR can be readily calculated from lift [18]. RR is a well-established standard measure of association strength that aligns with the causal framework. High RR values are indicative of a real association, making rules with high RR values attractive for further investigation. P(A) is highly correlated with the attributable excessive burden. Thresholding both RR and P(A) to reduce the number of rules for further investigation is consistent with the guiding principles in population health research.

It should be cautioned that the reduced rules obtained from the proposed process are exploratory and should be investigated carefully using a standard epidemiological framework such as multivariable analyses to control for potential confounding effects. For example, in the study of the adverse birth outcome, maternal risk factors should be adjusted. Multiple testing of a large number of hypotheses simultaneously in the data mining process can generate false-positive associations. Approaches such as permutation to correct for p-values and control false discovery should be considered.

Acknowledgements We thank the Alberta Perinatal Health Program for the birth data, the DoMiNO team for creating a multiple-source dataset, and Graham Erickson for performing the association rule mining. Lastly, Zhe Lu helped format the manuscript.

Funding The DoMiNO study was funded by the CIHR-NSERC Collaborative Health Research Program (FRN: 127789). Dr. Yuan is supported by NSERC (RGPIN-2019-04862). The funding sources did not influence the study design, the collection, analysis and interpretation of the data nor the writing of the manuscript.

References

1. Alberta Agriculture and Forestry. Alberta Climate Information Service (ACIS) [Available from: <http://agriculture.alberta.ca/acis/>]
2. Alberta Health Services. Alberta Perinatal Health Program [Available from: <https://aphp.dapasoft.com/Lists/HTMLPages/NewLandingPage.aspx>]
3. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*. **26**(2), 255–264 (1997)
4. DoMiNO Project. Data Mining & Neonatal Outcomes - Spatial Data Mining Exploring Co-Location of Adverse Birth Outcomes and Environmental Variables Project [Available from: <https://sites.google.com/a/ualberta.ca/domino/>]
5. Gordis, L.: *Epidemiology* 5th Edition: Saunders (2013)
6. Government of Canada. National Pollutant Release Inventory [Available from: <https://www.canada.ca/en/services/environment/pollution-waste-management/national-pollutant-release-inventory.html>]
7. Ha, S., Hu, H., Roussos-Ross, D., Haidong, K., Roth, J., Xu, X.: The effects of air pollution on adverse birth outcomes. *Environ. Res.* **134**, 198–204 (2014)
8. Hämäläinen W. Kingfisher—an efficient tool for searching for statistical dependency rules [Available from: <http://www.cs.joensuu.fi/~whamalai/kingfisher.html>]
9. Hämäläinen, W.: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowl. Inf. Syst.* **32**(2), 383–414 (2012)
10. Hernán MA, Robins, J.M.: *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC 2019. Available from: <https://www.hsph.harvard.edu/miguel-herman/causal-inference-book/>
11. Hill, A.B.: The environment and disease: association or causation? Sage Publications, 295–300 (1965)
12. Höfler, M.: Causal inference based on counterfactuals. *BMC Med. Res. Methodol.* **5**(1), 28 (2005)
13. Jalali-Heravi, M., Zaïane, O.R.: A study on interestingness measures for associative classifiers. *Proceedings of the 2010 ACM Symposium on Applied Computing*; Sierre, Switzerland. 1774306: ACM (2010). pp. 1039–1046
14. Li, J., Adilmagambetov, A., Jabbar, M.S.M., Zaïane, O.R., Osornio-Vargas, A., Wine, O.: On discovering co-location patterns in datasets: a case study of pollutants and child cancers. *GeoInformatica* **20**(4), 651–692 (2016)
15. Li, J., Zaïane, O.R., Osornio-Vargas, A.: Discovering statistically significant co-location rules in datasets with extended spatial objects. *International Conference on Data Warehousing and Knowledge Discovery*. Springer (2014)
16. Shazan, M., Jabbar, M., Zaïane, O.R., Osornio-Vargas, A.: Discovering spatial contrast and common sets with statistically significant co-location patterns. *Proceedings of the Symposium on Applied Computing*, ACM (2017)
17. Silverstein, C., Brin, S., Motwani, R.: Beyond market baskets: generalizing association rules to dependence rules. *Data Mining Knowl. Disc.* **2**(1), 39–68 (1998)

18. Vu, K., Clark, R.A., Bellinger, C., Erickson, G., Osornio-Vargas, A., Zaiane, O.R., Yuan, Y.: The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies. *BMC Med. Inf. Decision Making*. **19**(1), 112 (2019)
19. World Health Organization. How to interpret epidemiological associations. Available from: https://www.who.int/water_sanitation_health/dwq/nutrientschap9.pdf

Mathematics and Computation in Finance, Economics, and Social Sciences

About the Algorithms of Strategic Management



Manana Chumburidze, Mzia Kiknadze, Nino Topuria, and Elza Bitsadze

Abstract This article is devoted to the development of generalized dynamical programming methods of mathematical and computational approaches for solving optimization problems in modern business. The risk mitigation strategies of projects selection in corporate network of company have been discussed. The multistage graph-model of iterative planning projects has been constructed. The algorithm to solve optimization problem of project management with a minimized risk criteria has been delivered. The tools applied in this development based on the graph theory applications and queuing implementations.

Keywords Optimization problems · Graph theory · Dynamical programming

1 Introduction

Optimal management strategies in modern business is a base source on the latest innovations in enhancing main management functions, such as projects development in corporate network of company. Project management plays a crucial role in enabling companies to transform business and execute strategy effectively. Strategic project management that same level of structure and consideration can also be applied to selecting projects for organizations. Not every project is a good idea. Project selection methods [3, 7, 8] essential for an effective business plan. There are variety of documented methods for selecting project. There are a number of approaches

M. Chumburidze (✉) · M. Kiknadze · N. Topuria · E. Bitsadze
59str.Tamar Mepfe, Kuataisi, Georgia
e-mail: manana.chumburidze@atsu.edu.ge

M. Kiknadze
e-mail: mziakiknadze@gmail.com

N. Topuria
e-mail: nintopuria@gmail.com

E. Bitsadze
e-mail: elizabethsadze@mail.rue

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_59

to organizing and completing success of any project. Regardless of the methodology employed, careful consideration must be given to the overall project objectives, timeliness, risks and cost. There are several project management techniques through a focus on outcomes (benefits) of a project. This can help to reduce the risk of a completed project being a failure.

This paper describes a new technique forward and backward approaches of dynamical programming method (**DP**) [4, 5] which modifies the usual backtracking procedures of optimal policies. In our investigation this problem with generalized **DP** approaches as a powerful algorithms has been solved. In this paper optimization problem of projects management divides into the simple sequences of optimal project selection problems in which they are interrelated leading to decisions. Several classes of graph optimization problems, which can be solved using **DP**, are known to have more efficient tailor-made algorithms. In the **DP**, there is no standard formula that can be used to make a certain formulation. The success of **DP** lies in the fact that an optimum solution to a sub problem usually depends only on the optimum values of adjacent sub problems and not on the structure of these adjacent sub problems. In this paper we are using the multistage graph modeling method [1, 2, 11] because graph is a particular way for visualization of the storing and organizing dates of the investments and corresponding profits. In graph terminology, we may consider an elementary sub problem as a node and the computational effort to solve the sub problem as weight. The interrelationship between the sub problems is the underlying constraint graph showing which sub problems (nodes) can be combined. The final goal is to successively cluster all the nodes of the underlying constraint graph into one node.

2 Notations and Definitive Concepts

This paper discusses forward approaches (from **start** stage to **end** stage) to construct multistage graph for modeling of selected project investment and backward approaches (from **end** stage to **start** stage) to find path with minimal risk value and to know whether they make the same final decision [5]. The project under consideration is assumed to involve invest incurred over a period with a risk value. One of project management's primary functions is to mapping out a clear plan of the projects selection from the beginning to the end period of consideration [6, 10].

There are the following sub-problems discussed:

- Investigation dynamic model of multistage plan of project selection in corporate network company;
- Investigation method of generalized **DP** approaches;
- Justify an algorithms of solution;
- Create a pseudo code for software implementation;

Decision problem include the following stages:

- Construct the graph-model into several sequential stages of projects selection starting on from the stage **start** to the **end** stage. Where stages designates of current project selection. The vertex of graph are used for storing the sum invests and edges of graphs are labeled by the probability-risk of the corresponding project selection stage. The results obtained at each stage are used for the states in the next stage so that at the forward stage is obtained and used as a consideration of the decision in the next stage.
- Find out solution of optimization problem into several successive sequential stages starting from the stage-**end** by backward **DP** and interconnected with a decision rule in each stage.

Find the optimal solution with cost at next stage based on the characteristics of the **DP**, the case is divided into several stages and the decision has to be made at each stage. In the backward, used firstly stage is obtained and used as a consideration of the decision in the next stage. Cost backward always increase steadily, because the cost in the next stage depends on the cost in the previous stage and formed the decision of each stage by taking the smallest value previous sub-problem. Therefore backward approaches have the optimal result.

2.1 Statement Problem

In this section dynamic model of multistage plan of project selection in corporate network company has been investigated by **DP** approaches. Let us **m** is number of given application of projects in the same company. Two different projects cannot be overlapped in time. In every application. The start and finish time of project is indicated. Different applications can be but only one of them will be.

Introduce the following notations: $C = (C_{ij})_{n \times m}$ -matrix of initial values of projects investment in company, $R = (R_{ij})_{n \times m}$ -matrix of values of corresponding risks, $P = (P_{ij})_{n \times m}$ -matrix of projects, i-number of stage (period project selection); j-number of project in current stage (i=1,...,n;- j=1,...,m), **S**- total sum of projects invests, n-number of selection period of project, m-total number of projects within consideration, $X = (X_{ij})_{n \times m}$ -Boolean matrix [9] described of selection projects.

Let us consider mathematical formulation of weighted project scheduling problem(**WPS**):

WPS problem. It is required to find the minimum risk subset of projects such that no two projects in the subset overlap under the following conditions:

Constraint condition:

$$\sum_{i=1}^n \sum_{j=1}^m (C_{ij}[data]X_{ji}) \leq S$$

Criteria of optimization:

$$\sum_{i=1}^n \sum_{j=1}^m (R_{ij}[data]X_{ji}) \rightarrow \min$$

where data is variable of invest

WPS a special case of single-source shortest paths problems in graphs theory and it has the optimal substructure, so it can be solved by **DP** method.

2.2 Solution Problem

Let us $G = (V, E)$ -graph, where V is set of vertices, E is set of edges.

Definition. $G_\pi = (V_\pi, E_\pi)$ is a predecessor sub graph to $G=(V, E)$ with a source vertex- A , where

$$V_\pi = \{u \in V : \pi[u] \neq NILL\} \cup \{A\}; E_\pi = \{(\pi[u], u) : u \in V_\pi - \{A\}\}$$

where $\pi[u]$ -parent of u -node

Graph-modeling algorithm (**GMA**) from the **start** point to the **end** point is in a breadth ward motion and uses a data structure **queue** to remember to get the next vertex as start a new stage of invest. **GMA** used to model relations between of stages periods of selection projects to find the minimum risk spanning tree by dynamic programming approach. In particularly graph-creating algorithm include following sub stages:

- In first stage create a set of edges of adjacent vertices of start vertex and arrange them in order of invest and weight-age by correspond risk. we shall keep sum of invests from beginning to current stage in vertices:

$u[i][j].data = currentv.data + c[i][j]$ where v is vertex.

- Next we start adding edges of adjacent vertices to the graph beginning throughout from each one vertex. Correspond set of selection projects for last period satisfy following condition:

$$S - v[i][j] \geq 0$$

and continue until **end** (end vertex)stage. We will get to graph-data model of **WPS** problem (see **Fig:1**).

Let us consider pseudocode [4] of corresponding algorithm:

GMA(G, startv)

```
{
for  $\forall u \in V[G] \setminus \{startv, endv\}$ 
{ $u.data = 0; \pi[u] = NILL;$ }
 $Startv.data = 0; \pi[Startv] = NILL; Q = \emptyset;$ 
 $ENQUEUE(Q, startv);$ 
while (! $Q.empty()$ )
{
 $currentv =$ 
 $DEQUEUE(Q)$  for ( $i = 0; i < project.vertices.count - 1; i ++$ )
{ $u[i].data = currentv.data + c[i];$ 
if ( $S - u[i].data \geq 0$ )
{
```

```

(currentvertex, u[i]).Label = r[i];
add u[i] in adj.set[current-vertex];
if u[i] en-visited
ENQUEUE (Q,u[i]); }; };
};
}
    
```

After completing graph-modeling algorithm we use backward **DP** approaches from **end** stage to **start** stage to find path with minimal-risk and to get the final result. The algorithm treats the **end** vertex as a single source and apply the edge relaxation to the graph to obtain the minimal paths (minimal value of risk) to the adjacent vertices. It is possible to reconstruct the paths by repeatedly using the edge of relaxation. In this algorithm the risk-probability of the shortest path with the shortest-length route will calculated by save the predecessors for each vertex. See pseudo code of relaxation algorithm:

```

Relax(u, v, r)
{
if (d [v] > d[u] + r(u, v))
{d [v] = d[u] + r(u, v);
π [v] = u; }
}
    
```

where $d[v]$ is label of vertex- v .

A given problems has an optimal substructure property and **DP** algorithm is usefulness to solve it [10].

Lemma. Let us $G = (V, E)$ is weighted graph with weigh function $w : E \rightarrow R$, then after relaxation of edge $(u, v) \in E$ following inequality will be satisfied: $d[v] \leq d[u] + w(u, v)$

Proof. In case of a condition: $d[v] > d[u] + w(u, v)$, after relaxation of edge (u,v) we will get: $d[v] = d[u] + w(u, v)$, else if $d[v] < d[u] + w(u, v)$ then after relaxation of edge (u,v) the values of $d[v]$ and $d[u]$ do not changed, so after relaxation of edge (u,v) will be satisfied the following inequality: $d[v] \leq d[u] + w(u, v)$

Lemma. Let us $G = (V, E)$ is weighted graph with weigh function $w : E \rightarrow R$, if $p = \langle v_1, v_2, ..v_k \rangle$ is shortest path from v_1 to v_k and $1 \leq i \leq j \leq k$ then $p_{ij} = \langle v_i, v_{i+1}, ..v_j \rangle$ is a shortest path from v_i to v_j

Proof. If we break p -path up into parts: $v_1 \underbrace{p_{1i}} v_i \underbrace{p_{ij}} v_j \underbrace{p_{jk}} v_k$ then it will be satisfied:

$$w(p) = w(p_{1i}) + w(p_{ij}) + w(p_{jk}).$$

Let us exist a path from v_i to v_j , with a condition:

$$w(p'_{ij}) \leq w(p_{ij})$$

then $v_1 \underbrace{p_{1i}} v_i \underbrace{p'_{ij}} v_j \underbrace{p_{jk}} v_k$ is a path from v_1 to v_k with a weight:

$$w(p') = w(p_{1i}) + w(p'_{ij}) + w(p_{jk}) \text{ less than } w(p) \text{ but it is impossible because } p$$

is shortest path from v_1 to v_k

See pseudo code of **backward DP** algorithm:

Initialize-Single-Source(G, endv)

```
{
  for( $i = 1; i \leq V[G] - 1; i++$ )
  {for $\forall(u, v) \in E[G]$ 
Relax( $u, v, w$ ); }
};
```

After completing solution project selection' plan we should to print optimal plan of selection projects.

See forward approach algorithm to print a shortest path in graph described projects selection plan with minimization of risk:

```
PRINT - PATH( $G, \text{Startv}, v$ ) {
  print s;
  if ( $\pi [v] = \text{NIL}$ )
  print "path not found";
  else
  PRINT - PATH( $G, \text{Startv}, \pi [v]$ );
  print v; }
```

For simplest case, we have considered the particular example, when $n=3$, $m=9$ and $S=10$.

Initial data value of projects costs with corresponding risk are presented in **Table 1**. Let us consider one of them stages of forward approaches to create graph-model (see **Fig. 1**):

- We start from visiting **start** vertex and mark it as **0**.
- In the next we create adjacent-list of **start** vertex and save the initial dates of project cost and label of edges by the corresponding risks. Accordingly **Table 1**. With respect to **start** vertex **0** we have three adjacent node (three alternative projects).
- In order to invest we choose first and mark it as **1** and create label of edge by **0,6**, then put it in queue (**Q**).
- In the next, we choose second and mark it as **2** and create label of edge
- In the next, we choose third and mark it as **3** and create label of edge by **0,9**.
- In the next, put it in **Q**.
- Now, **0** is left with no en-visited adjacent nodes. So, we remove from **Q** and find **1**. After repetitively perform of similarly stages we will get the graph data model of **WPS**.

allow us consider the **stages of backward approaches to get minimal risk value**:

- We are starting from visiting **10** (**end** vertex) and perform relaxation of adjacent vertices in order **5, 6, 8** creating corresponding minimal risks labels: **0,9; 0,7; 0,5**.
- In the next starting from visiting **5** to perform relaxation of adjacent vertex **2** by creating corresponding minimal risk's label **1,6**.
- In the next starting from visiting **6** to perform relaxation of adjacent vertices in order **1,3** by creating corresponding minimal risks labels **1,5; 1,4**.

Table 1 Initial dates of WPS

Project[i][j]	First company	Second company	Third company
Select period	C[i][j] R[i][j]	C[i][j] R[i][j]	C[i][j] R[i][j]
First period	1 0,6	3 0,7	2 0,5
Second period	2 0,8	5 0,8	4 0,7
Third period	3 0,9	6 0,9	5 0,9

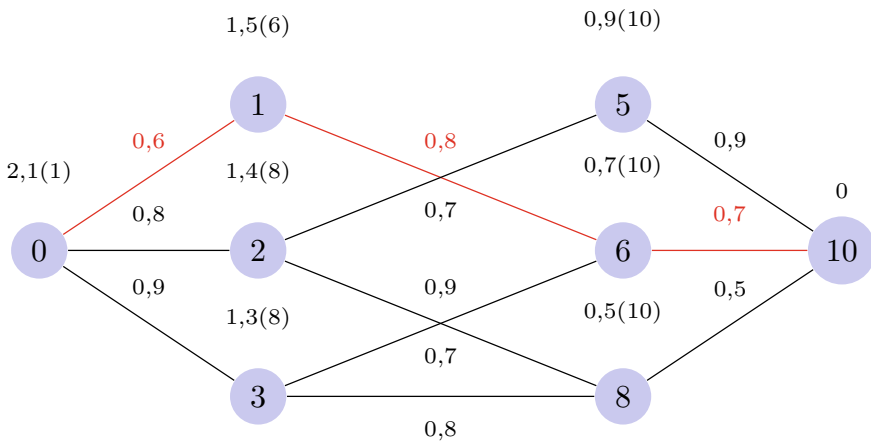


Fig. 1 Graph of projects selection

- In the next starting from visiting **8** to perform relaxation of adjacent vertices in order **2,3** by creating corresponding minimal risks labels **1,4; 1,3**.
- In the next starting from visiting **1** to perform relaxation of adjacent vertex **0** by creating corresponding minimal risk's label **2,1**;
- In the next starting from visiting **2** to perform relaxation of adjacent vertex **0** by creating corresponding minimal risk's label **2,1**;
- In the next starting from visiting **3** to perform relaxation of adjacent vertex **0** by creating corresponding minimal risk's labels **2,1**. After completing minimal risk finding algorithm, we will get the project selection plan and we perform of algorithm to print up of path selection plan.

Best plan of project scheduling problem has a following result: For first company will be selected project of first period with risk **0,6**; For second company will be selected project of second period with risk **0,8**; For third company will be selected project of second period with risk **0,7**. Selection projects will get the minimal risk value **2,1**. The result of selection projects in the red line is exposed (see Fig. 1).

3 Result and Discussion

In this work optimization problem of project management by the generalized **DP** approaches to minimized projects risk has been solved. The particular example of project selection problem within corporate network of company has been discussed. Multistage graph-model to describe of projects invest with a corresponding risk-probability has been constructed. The algorithms to build risk mitigation plan related to minimal path in graph have been delivered. The planning results to all selected project for implementation have been considered.

This investigation has a many advantages: forecasting and analyzing of projects risk in any time; fluently make decision in the planning stage to select of candidates; monitoring and managing projects flows; optimize projects flow; efficient in terms of time complexity; the results enable to be applied in decision making problems to optimize solution.

References

1. Chumburidze, M., Basheleishvili, I., Khetsuriani, A.: Dynamic programming and greedy algorithm strategy for solving several classes of graph optimization problems. *BRAIN* **10**(1), 101–107 (2019)
2. Chumburidze, M., Basheleishvili, I.: The complexity of algorithms for optimization problems. *Comput. Sci. Telecommun.* **54**(2), 125–130 (2018)
3. Chumburidze, M., Lekveishvili, D.: Numerical approximation of basic boundary-contact problems. *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (2017). <https://doi.org/10.1115/DETC2017-67097>
4. Bennett, N.: Introduction to algorithms and pseudocode, exploring modeling and computation (2015). <https://doi.org/10.13140/RG.2.2.28657.28008>
5. Wang, L. et al.: A hybrid backtracking search optimization algorithm with differential evolution. *Hindawi Publis. Corporat. Mathemat. Probl. Eng.* **2015**(769245), 16 (2015)
6. Gogichaishvili, G., Surguladze, G., Topuria, N., Urushadze, B.: Construction of management information systems of distributed business processes based on petri networks and object-role modeling. *Bull. Georg. Nat. Acad. Sci.* **8**, 58–64 (2014)
7. Kononenko, I., Kharazii, A.: The methods of selection of the project management methodology. *Int. J. Comput.* **13**(1), 8 (2014)
8. Elsayed, S., Sarker, R., Essam, D.: Adaptive configuration of evolutionary algorithms for constrained optimization. *Appl. Mathemat. Comput.* **222**, 680–711 (2013)
9. Gudder, S., Latrémolière, F.: Boolean inner-product spaces and boolean matrices. *Linear Alge. Appl.* **431**, 274–96 (2009)
10. Chapman, C., et al.: Selecting an approach to project time and cost planning. *Int. J. Proj. Manag.* **3**(1), 19–26 (1985)
11. Bondy, J., Chvatal, V.: A method in graph theory. *Disc. Mathemat.* **15**(2), 111-135 (1976)

Using Cognitive Fit Theory to Evaluate the Effectiveness of Financial Information Visualization: An Example Using Data to Detect Fraudulent Transactions



A. Czeglédi, L. Scott Campbell, and D. Smiderle

Abstract The objective of this research was to investigate the impact that financial data visualization (DV) has on decision making in detecting fraudulent transactions. This study was focused on the effects of financial DV formats on accuracy and speed. According to the Cognitive Fit Model (CFM) the effectiveness of the problem-solving process is a function of the relation between the problem-solving task and problem representation. Participants of research study (95 accounting undergraduate students) were assigned to different groups, each group was presented with the same financial information in different formats: text, table and DV; and asked to identify potentially fraudulent transactions. The study results suggested a strong relation between presentation format and speed for decision making. This result could have practical application: in order to enhance the decision making, organizations could consider the presentation format of their financial data if decision is time sensitive.

Keywords Business information visualization · Information visualization · Cognitive fit · Fraud detection

1 Introduction

The main objective of this research was investigation of impact of data visualization (DV) of financial information on decision making for business related scenarios. Accountants are required to analyze data, use professional judgment and to see “the big picture” in order to provide the right information at the right time for making

A. Czeglédi (✉) · L. Scott Campbell
Conestoga College, Kitchener, Canada
e-mail: Aczeglédi@conestogac.on.ca

L. Scott Campbell
e-mail: lscottcampbell@conestogac.on.ca

D. Smiderle
Principal, Scentre Educational Services Inc., Guelph, Canada
e-mail: davesmiderle1@gmail.com

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_60

financial decisions [12, 14]. Accountants add greater value to their organizations by spending more time analyzing, interpreting, and providing information for decision making rather than time spent preparing standardized reports [6, 15, 25].

In the current business environment professional accountants are required to provide effective and accurate decision support based on large, complex data sets. It takes significant time, effort and cost to understand and analyze information; and it is even more challenging to present financial information in a concise and effective format understandable by different levels of stakeholders. Accountants are looking for techniques to interact with complex financial data and thoughtfully crafted visualization can increase our understanding of data, it could help to see a big picture: reveal patterns, quantities, changes over time, or recurring themes at a glance [26]. Engin and Vetschera [10] emphasized importance of appropriate information presentation form. Inappropriate presentation form may lead to weaker performance in decision making for current and subsequent tasks.

Professional accountants are required to use effective and efficient methods to communicate complex financial information in a timely manner. One possible approach to achieve this task is to adopt DV. DV could help accountants with analyzing financial data and deal with large data volume, diversity and complexity of information [14, 16]. Today evolving technologies and increasing streams of data create the need for visualizing effects, possibilities, and consequences. DV could help to reduce the information volume to a manageable size, could help to focus on crucial data points, to gain insights, draw conclusions by formulating theories on the basis of patterns, themes and calculations, and present and convey an effective message to the stakeholders [2, 14, 26].

DV converts data to a visual, user friendly format by creating a story [14, 26] about business performance with visual impact that demonstrates interconnections between various elements of operations and results. The messages in these stories could leave a longer lasting impression on users of financial information compared to standardized reports. DV has been already successfully utilized in many fields, for example, in medicine, genetics, biology, forest planning, engineering, insurance to name a few, and it is safe to predict that additional applications of DV will be discovered in the future. The main goals of DV are to explore and explain the data and present information in a format that engages the human's cognitive and visual abilities [2]. DV helps to understand complex information by accumulating, grouping, and displaying information in a more effective way [9].

Dilla, Janvrin, and Raschke [7, p. 1] defined DV as "computer-supported visual representation of data that allows users to select the information they wish to view and its format". Data visualization enables easy communication and understanding of large and complicated data sets. It enables the users to get a better appreciation and comparison of the effects of the data [13]. The main purpose of data visualization is to visually illustrate or communicate data and or information in a manner where readers could easily comprehend the information without much need for quantitative or qualitative support and see into clear patterns and view the complex patterns or relationships uncovered in the data mining process [28, 31]. Data visualization assists with identifying trends, patterns, outliers, and or correlations between variables. It also

enables the exploration of relationship that could be embedded in statistical models and is fundamental to readers' comprehension, interpretation, and understanding of the information presented [4]. Rodriguez and Kaczmarek [26] highlighted few functions of DV: making comparison (could save time for analysis), establishing connections (could help to understand interconnection among data), drawing conclusions and help to solve complex questions.

There are two elements of DV that could impact the decision-making process in an accounting environment. The first one is interactivity, or ability of user to manipulate information views during the decision-making process. The second element of DV is visualization or "the manner in which data are depicted or portrayed" [1, 7, 19]. Characteristics of DV could impact the decision-making processes and outcomes by changing the decision-making frame of reference, which information to use and how. DV could improve the decision-making process by providing ability to select, navigate, and restructure complex data [19], although DV may also lead to overconfidence and biases in decision-making. Dull and Tegarden [9] explained that from a cognitive science perspective, DV can improve problem solving capabilities, as described by Miller [20] that a human's input channel capacity can be increased by using visual abilities and reduce information overload. Schkade and Kleinmuntz [27] noted significant impact of information presentation format on the decision process. Effective DV should include the following characteristics [26]: to be universal, immediate, concise, inviting, memorable, revealing, reusable and versatile.

In business environment multiple types of DV are currently in use: statistical software packages, technical computing languages, visualization software packages, open source programming languages, close-ended web-based libraries, open web-based libraries, and custom code in proprietary programming language [4, 11, 26]. These solutions range in programming skills required, design flexibility, and interoperability.

Type of DV tools for decision making depends on application, and rather than recommending the use of specific DV. Rodriguez and Kaczmarek [26] suggested getting familiar with framework in order to select appropriate type of DV tools in each situation. Several metrics should be considered, such as size of project, type of users (internal or external), available resources and capabilities.

This study focused on information presentation effects on a specific type of management judgment or decision making—identification of potentially fraudulent transactions. Fraud identification was selected due to critical importance of this issue [8]. Organizations are looking for ways to combat fraud; an estimated \$3.5 trillion in revenue annually is lost due to fraud and it is recommended that DV could be applied to complex fraud challenges [2, 14].

2 Research Question

The Cognitive Fit Model has been used in developing research question and the hypothesis for this study. According to the Cognitive Fit Model (CFM) the effectiveness of the problem-solving process is a function of the relation between the problem-solving task and problem representation [32, 33]. Vessey [32, p. 223] concluded that “matching the problem representation to the type of task to be solved results in improved decision-making performance“. Dull and Tegarden [9] emphasized that a mismatch between the problem representation and the problem-solving task could lead to a reduction in speed and accuracy in the decision-making process. Vessey and Galletta [33] noted that based on information processing theory, human problem solvers will seek ways to reduce their problem-solving effort, since they are limited information processors [22]. Reduction in problem solving effort could be achieved by matching the problem representation to the task, an approach that is known as Cognitive Fit [32]. CFM is a cost–benefit characteristic that suggests that for the most effective and efficient problem solving to occur, the problem representation and any tools or aids employed should all support the strategies (methods or processes) required to perform that task [32]. This means that the problem representation a problem solver uses must be considered in the context of the task to be solved. Based on CFM, it could be expected that there should be a difference in decision making performance based on the DV format used for experiment. Similar to Dull and Tegarden [9] approach it could be expected that the better “fit“ of DV the more accurate the prediction of participants. When assessing decision making process, we should also consider the time spent to make a decision. Vessey [32] believes there are only two objective performance variables for decisions: decision time and accuracy. As such, our research question and hypotheses are as follow:

RQ: This study investigated the possible impact of the data visualization formats on two variables speed and accuracy in business scenarios decision making process, specifically related to the potentially fraudulent transactions. Therefore, our H0 hypotheses are as following:

H0.1: *Employing data visualization has no impact on the speed of the task in business scenarios decision making process.*

H0.2: *Employing data visualization has no impact on the accuracy in business scenarios decision making process.*

Our H1 hypotheses are as following:

H1.1: *Employing data visualization has impact on the speed of the task in business scenarios decision making process.*

For first hypothesis our goal was to investigate if participants presented with data visualization presentation format will able to complete task with higher speed (less time) compared to control group with text format (we also included group with table format presentation in our experiment).

H1.2: *Employing data visualization has impact on the accuracy in business scenarios decision making process.*

For second hypothesis our goal was to investigate if participants presented with data visualization presentation format will able to complete task with higher accuracy (less errors) compared to control group with text format (we also included group with table format presentation in our experiment).

The Independent Variables of this research project are same financial information presented to participants in three different formats: text format as a text document (neutral “control”), tables format, and in DV info graphic format (easily visualized). Currently the most common technique for visualizing financial data remains the standard charts (e.g., line, bar, pie charts, and their variations [16, 21]. DV can facilitate the analysis of data by improving business dimension of information, and help decision makers in finding trends and key events [30].

The Dependent Variables. This research experiment included two dependent variables: accuracy and speed. Accuracy and speed were selected as Bačić and Fadlalla [3] had considered speed of task, accuracy and recall metrics as fundamental metrics for decision making in business contexts.

3 Research Approach

The experiment was designed to investigate the impact of DV on the outcome measures mentioned above. An experiment was conducted to investigate the hypothesis and research question. Below we described methods to develop the experiment and business scenarios. Research data was collected from experiments completed for this study. This experiment utilized a Qualtrics application for data collection and allowed the researchers to obtain all required information, including measurement of time to assess all business scenarios by participants. Pre-testing was completed prior to completing experiment on a larger population.

3.1 Participants

Participants for this research were accounting students from same undergraduate program, which allowed for control of financial knowledge capabilities levels of the participants, in addition to pre-experimental training provided to all participants. Although accounting students are not currently business decision makers, they are future managers and similar to actual managers in terms of decision biases [9, 10]. Participants were contacted in their classes, where faculty agreed to conduct this experiment. All experiments were conducted at the lab under control environment. A reward system was developed to attract and compensate the participants (1% extra credit in their courses for participation in this study).

Participants were randomly assigned to different groups and depending on group assigned asked to analyze financial information in a specific format. Three groups were utilized with each group presented with the same financial information in

different formats: text format, table format and DV format. Each group was asked to review five business scenarios to assess their speed of task and accuracy.

3.2 Measures

Dull and Tegarden [9] in their study of comparison of three visual representations had focused on prediction on wealth based on different presentation formats (two-dimensional line graph, modified trajectory line graph, modified trajectory line graph with ability to rotate image). They used overall wealth prediction (accuracy) and time as dependent variables. We utilized their approach in measuring our dependent variables. The data collected from participants included serial numbers of transaction which possibly include fraudulent transaction and require future investigation. For each set of business scenario participants found possible fraudulent transactions. The dependent variable is accuracy or number of correct potentially fraudulent items identified by participants (fraudulent items accuracy—FIA) based on presented information, measured for each participant, for all business scenarios, by subtracting the participant's unidentified potentially fraudulent items (PFUI) from total number of fraudulent items in each business scenario/model (MFI).

$$\text{FIA} = \text{MFI} - \text{PFUI} \quad (1)$$

Potentially fraudulent items for business scenarios were created by researchers based on their professional experience in fraudulent transaction detection in industry and were reviewed by other professionals from industry.

A second dependent variable was speed or the time it takes participant to assess all business scenarios and make decision about potentially fraudulent item. It was measured as the number of seconds to assess all business scenarios. There was no fixed time limit, although participants encouraged to complete experiment in 30 min.

3.3 Procedure

Experiment for this study included two steps: first participants were provided an overview of possible signs of fraudulent activities, second—participants assessed five business scenarios to identify potentially fraudulent activities based on three types of financial data presentation format (one type per group) groups were randomly selected (see Fig. 1). Our experiment design was based on modification of the Solomon four-group design [29]. Participants were randomly assigned to groups and completed training in order to get understanding how to identify potentially fraudulent transactions same as presented in all five business scenarios. Each business scenario included 30 items/transactions for each format of financial information presentation—text, table and DV.

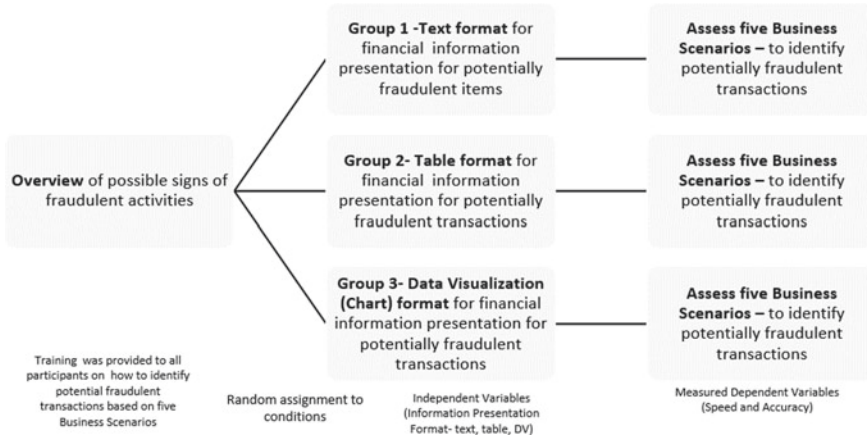


Fig. 1 Experiment design

The data collected from participants included serial numbers of transaction which include possibly fraudulent items and require future investigation. Based on nature of data for our experiment (fraud investigation, not complex data), potential users of DV (accounting students), and available resources we had selected the following DV format: prebuilt solutions that required no programming (visualization software packages).

3.4 Business Scenarios Design

There are two elements of DV that could impact the decision-making process in an accounting environment: interactivity and visualization, our research focused on the second, the “presentation of data” [1, 7, 19]. Data to create a business scenario was based on analysis of a data set using ACL Analytics (formerly known as Audit Command Language) software. The participants reviewed information to detect transaction anomalies, which is an important fraud detection procedure. Effective data visualization has the potential for making the detection of fraudulent transactions more efficient and effective. Currently research of effectiveness of data visualization for fraud detection is limited [8]. DV could help to convert information to a manageable scope, identify and prioritize threats, develop critical intelligence and make effective decisions. DV can create the story behind the data, and in cases of potential fraud, can demonstrate linkages that are not obvious between people, places and financial and non-financial potentially fraudulent information [2, 14].

According to PWC 2018 Global Economic Crime and Fraud Survey [24], an asset misappropriation, consumer fraud and cybercrime were the most frequently reported frauds across industries. For financial services an asset misappropriation and consumer fraud are the most common fraud organizations had experienced recently.

Based on this information we had focused on consumer and asset misappropriation types of fraud in designing our business scenarios. Application of visualization techniques for fraud detection could help to recognize and present data anomalies, which could make the identification and quantification of fraud schemes much easier [23]. Engin and Vetschera [10] recommended to use different types of problems to enhance information representation in a decision-making experiment. Scenarios were selected that closely reflect conditions, which decision makers are facing in practice as recommended by Bačić and Fadlalla [3]. They had noted that in past research task types are too abstract or do not effectively reflect decision making tasks in practice. Benford's law was used in designing one of our business scenarios, as application of this law to a population of transactions is common practice in fraud detection investigations. [17]. Benford's law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small [5]. Participants reviewed data for adherence with Benford's law, enabling the fraud detection, as fraudsters are usually not familiar with this digital law and tend to invent numbers with approximately equal digital frequencies [17]. Participants, as part of each business scenario, required to assess the following:

- Business scenario 1—Identify annuity payments paid to the annuitants over certain age, for example 90 years old, where there has been no prior confirmation that annuitant is still alive (type of fraud—consumer fraud, payments to the annuitants who are not alive).
- Business scenario 2—Identify monthly payments over certain dollar limit, for example over \$10,000, where there has not been approval by two managers (type of fraud—assets misappropriation (AM), payments made without appropriate approval).
- Business scenario 3—Identify annuity contracts where interest rates at the time of issue were higher than certain amount, for example 2% (type of fraud—AM, incorrect interest rate).
- Business scenario 4—Identify annuity contracts that warrant further investigation based on Benford's analysis of annuity payment amounts and contract principal balances (type of fraud—AM, fictitious contract for nonexistent customer).
- Business scenario 5—Identify annuity contracts where no agent's information attached or inactive agent (type of fraud—AM, fictitious contract for nonexistent customer).

Each business scenario included 30 items/transactions, table 1 includes short extract of business scenario 1 (only five transactions/items were included for each presentation type consider size limitation for this publication).

Table 1 Example of business scenario

Group	Business Scenario 1 —Your supervisor has requested your assistance in a review currently under way. Your team is reviewing 5000 active annuity contracts for potential fraudulent transactions related to monthly annuity payments. Monthly annuity payments should be discontinued for annuitants older than age 90 unless a confirmation letter has been received from the annuitant. She has asked you to review a batch of 30 records. She would like you to identify accounts that might require further investigation																																																						
Group 1-Text	<ol style="list-style-type: none"> 1. Annuity contract 6,340,477 provides monthly annuity payments for an individual who is currently 71 years old. A confirmation letter is not on file for this annuitant 2. Annuity contract 7,307,734 provides monthly annuity payments for an individual who is currently 98 years old. A confirmation letter is not on file for this annuitant 3. Annuity contract 2,923,019 provides monthly annuity payments for an individual who is currently 70 years old. A confirmation letter is not on file for this annuitant 4. Annuity contract 9,409,896 provides monthly annuity payments for an individual who is currently 69 years old. A confirmation letter is not on file for this annuitant 5. Annuity contract 6,995,474 provides monthly annuity payments for an individual who is currently 67 years old. A confirmation letter is not on file for this annuitant 																																																						
Group 2-Table	<table border="1"> <thead> <tr> <th>Record</th> <th>Account</th> <th>Prov</th> <th>Balance</th> <th>Account Created</th> <th>Agent</th> <th>DOB</th> <th>Age</th> <th>Confirmation Received</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>6340477</td> <td>ON</td> <td>\$ 275,325.32</td> <td>1-5-14</td> <td>10023</td> <td>1948-01-01</td> <td>71</td> <td></td> </tr> <tr> <td>2</td> <td>7307734</td> <td>AB</td> <td>\$ 324,635.00</td> <td>1-6-14</td> <td>10021</td> <td>1921-01-01</td> <td>98</td> <td>No</td> </tr> <tr> <td>3</td> <td>2923019</td> <td>ON</td> <td>\$ 407,774.00</td> <td>1-6-14</td> <td>10003</td> <td>1949-01-01</td> <td>70</td> <td></td> </tr> <tr> <td>4</td> <td>9409896</td> <td>QC</td> <td>\$ 292,466.00</td> <td>1-6-14</td> <td>10009</td> <td>1950-01-01</td> <td>69</td> <td></td> </tr> <tr> <td>5</td> <td>6995474</td> <td>ON</td> <td>\$ 273,889.78</td> <td>1-6-14</td> <td>10018</td> <td>1952-01-01</td> <td>67</td> <td></td> </tr> </tbody> </table>	Record	Account	Prov	Balance	Account Created	Agent	DOB	Age	Confirmation Received	1	6340477	ON	\$ 275,325.32	1-5-14	10023	1948-01-01	71		2	7307734	AB	\$ 324,635.00	1-6-14	10021	1921-01-01	98	No	3	2923019	ON	\$ 407,774.00	1-6-14	10003	1949-01-01	70		4	9409896	QC	\$ 292,466.00	1-6-14	10009	1950-01-01	69		5	6995474	ON	\$ 273,889.78	1-6-14	10018	1952-01-01	67	
Record	Account	Prov	Balance	Account Created	Agent	DOB	Age	Confirmation Received																																															
1	6340477	ON	\$ 275,325.32	1-5-14	10023	1948-01-01	71																																																
2	7307734	AB	\$ 324,635.00	1-6-14	10021	1921-01-01	98	No																																															
3	2923019	ON	\$ 407,774.00	1-6-14	10003	1949-01-01	70																																																
4	9409896	QC	\$ 292,466.00	1-6-14	10009	1950-01-01	69																																																
5	6995474	ON	\$ 273,889.78	1-6-14	10018	1952-01-01	67																																																
Group 3-DV	<p>A bar chart with the y-axis labeled 'Age' ranging from 50 to 120 in increments of 10. The x-axis is labeled 'Ages of annuitant/account holder' with categories 1 through 5. Five blue bars represent the ages of the annuitants: 71, 98, 70, 69, and 67. An orange diamond is placed above the first bar (71) to indicate that a confirmation letter is required for this annuitant.</p>																																																						

4 Results and Discussion

4.1 Data Reduction and Analysis Plan

102 Accounting students participated in the study. Using z score transformation to detect outliers the final sample size was reduced to 95 participants. Each group included about 32 participants.

With 2 dependent measures (Speed and Accuracy) and 1 independent measure (Presentation Format) MANOVA was used to analyze the data. Table 2 below provides the summary descriptive statistics for each group.

Table 2 Descriptive Statistics for all five business scenarios

1 = text, 2 = table, 3 = DV		Mean	Std. Deviation	N
Speed (in total seconds)	1.00	346.2109	111.4044	32
	2.00	289.0003	144.7628	32
	3.00	239.4119	136.0330	31
	Total	292.0898	137.2059	95
Accuracy (total number correct)	1.00	23.0000	2.0161	32
	2.00	22.2188	4.0140	32
	3.00	22.3548	3.7198	31
	Total	22.5263	3.3449	95

Table 3 Multivariate tests for five business scenarios

Effect		Value	F	Hypothesis df	Error df	Sig.
Presentation Format	Pillai's Trace	0.106	2.575	4.000	184.000	0.039

Table 4 Tests of between participants effects for all five business scenarios

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Speed	180,060.238	2	90,030.119	5.211	0.007
	Accuracy	11.119	2	5.559	0.492	0.613

4.2 Speed and Accuracy

Did varying the presentation format (text, table or data visualization) of the data have an impact on the speed or accuracy of our participants? The results of our study show a significant main effect of the Presentation Format on the dependent measures with a F value of 2.575, with a significant value of 0.039 (Table 3).

Further analysis as provided in Table 4 demonstrates that the Presentation Format had a significant impact on the Speed of the participants, with those who were given the Data Visualization format showing the quickest time ($p = 0.007$). There was no significant impact of Presentation Format on Accuracy ($p = 0.613$). All three groups show similar scores in regard to their ability to identify the correct answer to the business scenarios presented.

4.3 Discussion

Speed and accuracy are two critical elements in decision making [3]. The objective of this research was to investigate the impact that financial data visualization (DV) has on decision making in detecting fraudulent transactions. This study was focused on the

effects of financial DV formats on accuracy and speed. Participants of research study were assigned to different groups, each group was presented with the same financial information in different formats: text, table and DV; and asked to identify potentially fraudulent transactions in five business scenarios. The study results suggested a strong relation between presentation format and speed for decision making, decision makers presented with DV format had made decisions faster, with equivalent accuracy. This finding is consistent with Cognitive Fit theory [32, 33], that presentation format should be based on task.

The results of this study should be considered when reflecting on how to improve the efficiency of an Accountant's role in Industry. The study results suggested a strong relation between presentation format and speed for decision making. Decision makers presented with DV format had made decisions faster. This result is useful in those organizations whereby decisions around this type of data is time sensitive in some way. In addition, it could help business educators to focus on effective teaching formats. Further research is recommended on exploring exactly which conditions presenting information in a DV format would be beneficial.

Possible limitations of our study are: sample size, simplify problems for the experiment, student participants (convenience sampling) [18] and the possibility that prior knowledge of accounting principles may have negated the impact of presentation format. Future study could include additional variables such as memory retention (recall) and confidence [3] in decision making process.

Acknowledgements We appreciate everyone that supported and contributed to the completion of this project and paper, including CAAA (Canadian Academic Accounting Association), T.Basdeo, reviewers and editors.

References

1. Ajayi, O.: Interactive Data Visualization in Accounting Contexts: Impact on User Attitudes, Information Processing, and Decision Outcomes. Doctoral diss., University of Central Florida (2014). <https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=4013&context=etd>
2. Alawadhi, A.: The Application of Data Visualization in Auditing. Rutgers University-Graduate School-Newark, PhD diss. (2015)
3. Bačić, D., Fadlalla, A.: Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decis. Support Syst.* **89**, 77–86 (2016)
4. Bendoly, E.: Fit, bias, and enacted sense making in data visualization: frameworks for continuous development in operations and supply chain management analytics. *J. Bus. Logis.* **37**(1), 6–17 (2016). <https://doi.org/10.1111/jbl.12113>
5. Berger, A., Hill, T.P.: Benford's law strikes back: no simple explanation in sight for mathematical Gem. *Mathemat. Intell.* **1**(85) (2011)
6. Cory, S.N., Pruske, K.A.: A factor analysis of the skills necessary in accounting graduates. *J. Bus. Accoun.* **5**(1), 21–28 (2012)
7. Dilla, W., Janvrin, D.J., Raschke, R.: Interactive data visualization: new directions for accounting information systems research. *J. Inform. Syst.* **24**(Fall), 1–37 (2010)

8. Dilla, W., Raschke, R.: Data visualization for fraud detection: practice implications and a call for future research. *Int. J. Account. Inf. Syst.* **16**, 1–22 (2015)
9. Dull, R.B., Tegarden, D.P.: A comparison of three visual representations of complex multidimensional accounting information. *J. Inf. Syst.* **13**(2), 117 (1999)
10. Engin, A., Vetschera, R.: Information representation in decision making: the impact of cognitive style and depletion effects. *Decis. Support Syst.* **103**, 94–103 (2017)
11. Ertug, G., Gruber, M., Nyberg, A., Steensma, H.K.: From the editors—a brief primer on data visualization opportunities in management research. *Acad. Manag. J.* **61**(5), 1613–1625 (2018). <https://doi.org/10.5465/amj.2018.4005>
12. Gamage, P.: Big data: Are accounting educators ready? *Account. Manag. Inf. Syst. Cont. Si Informatica De Gestione* **15**(3), 588–604 (2016)
13. George, G., Osinga, E.C., Lavie, D., Scott, B.A.: Big data and data science methods for management research. *Acad. Manag. J.* **59**(5), 1493–1507 (2016). <https://doi.org/10.5465/amj.2016.4005>
14. Griffin, R.: Using big data to combat enterprise fraud. *Financ. Execut.* **28**(10), 44–47 (2012)
15. Institute of Management Accountants (IMA): Building a team to capitalize on the promise of big data (2016). <https://www.imanet.org/insights-and-trends/technology-enablement/building-a-team-to-capitalize-on-the-promise-of-big-data>
16. Ko, S., Cho, I., Afzal, S., Yau, C., Chae, J., Malik, A., Beck, K., Jang, Y., Ribarsky, W., Ebert, D.: A survey on visual analysis approaches for financial data. *Comput. Graph. Forum* **35**(3), 599–617 (2016)
17. Kossovsky, A.: Benford’s law: theory, the general law of relative quantities, and forensic fraud detection applications. World Scientific, New Jersey (2014)
18. Leedy, P.D., Ormrod, J.E.: Practical research. Planning and design. Pearson, Boston, MA (2019)
19. Lurie, N.H., Mason, C.H.: Visual representation: implications for decision making. *J. Market.* **71**(1), 160–177 (2007)
20. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81 (1956)
21. Murphy, J.J.: Technical analysis of the financial markets: a comprehensive guide to trading methods and applications. Penguin (1999)
22. Newell, A., Simon, H.A.: Human Problem Solving. Prentice-Hall, Englewood Cliffs, N.J. (1972)
23. Ngai, E.W.T., Yong, Hu., Wong, Y.H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decis. Support Syst.* **50**(January), 559–569 (2011). <https://doi.org/10.1016/j.dss.2010.08.006>
24. PWC 2018 Global Economic Crime and Fraud Survey. <https://www.pwc.com/gx/en/services/advisory/forensics/economic-crime-survey.html>. Retrieved March 26, 2019
25. Robert Half: 2016 Salary guide: accounting and finance (2016). Available online at: https://www.roberthalf.ca/sites/roberthalf.ca/files/rh-pdfs/atfamr_0915_iapdf_sg2016_can_eng.pdf
26. Rodriguez, J., Kaczmarek, P.: Visualizing financial data. Wiley, Indianapolis, IN (2016)
27. Schkade, D.A., Kleinmuntz, D.N.: Information displays and choice processes: differential effects of organization, form, and sequence. *Organ. Behav. Hum. Decis. Process.* **57**(3), 319–337 (1994)
28. Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E.: Knowledge management and data mining for marketing. *Decis. Support Syst.* **31**(1), 127–137 (2001)
29. Solomon, R.L.: An extension of control group design. *Psychol. Bull.* **46**(2), 137 (1949)
30. Sorenson, E., Brath, R.: Financial visualization case study: Correlating financial time series and discrete events to support investment decisions. *Information Visualisation (IV)*, 2013 17th International Conference, pp. 232–238. IEEE (2013)

31. Turban, E., Aronson, J.E., Liang, T.P., Sharda, R.: *Decision Support and Business Intelligence Systems*, Eighth ed, Pearson Education (2007)
32. Vessey, I.: Cognitive fit: a theory-based analysis of the graphs versus tables literature. *Decis. Sci.* **22**(2), 219–240 (1991)
33. Vessey, I., Galletta, D.: Cognitive fit: an empirical study of information acquisition. *Inf. Syst. Res.* **2**(1), 63–84 (1991)

Factors Affecting Sustainable Development and Modelling



Zurab Gasitashvili, Mzia Kiknadze, Taliko Zhvania, and David Kapanadze

Abstract A good access to information and intellectual resources and, consequently, their efficient management is important to operate an organizational system successfully. In any case, only based on the information processes can the entity make the decision. Based on the example of sustainable development of the region, the paper deals with the management of complex organizational systems and decision-making. To study the problems of regional development management, it is necessary to investigate the degree of influence that regional development factors (indicators) have on the criteria characterizing the regional development. The paper refers to the issue of selection of a subspace of basic factors of high importance (having a high degree of influence) from the space of factors of regional development. The study was carried out by the methods of perceptive-cognitive modeling, statistical analysis, fuzzy sets and graph theory.

Keywords Organization systems · Math models · Sustainable development

Z. Gasitashvili (✉) · M. Kiknadze · D. Kapanadze
Georgian Technical University, 77, Kostava str., Tbilisi 0175, Georgia
e-mail: zurgas@gtu.ge

M. Kiknadze
e-mail: mziakiknadze@gmail.com

D. Kapanadze
e-mail: david@gtu.ge

T. Zhvania
Department of Informatics, Guram Tavartkiladze Tbilisi Teaching University,
5, Samghereti str., Tbilisi 0101, Georgia
e-mail: talizhvania@gmail.com

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_61

1 Introduction

A good access to information and intellectual resources and, consequently, their efficient management is important to operate an organizational system successfully. In any case, only based on the information processes can the entity make the decision [1, 2].

During analyzing the complex organization systems, it is possible to identify a multitude of goals and factors that the system faces. Commonly, these goals are structured, so, they consist of sub-goals, each sub-goal may consist its own sub-goals, an so on. At the last stage, the list of atomic goals will be obtained. The number of sub-goals quantity may be very large. It should also be borne in mind that the individual elements of these atomic goals do not have the same primary purpose.

The problem of analyzing complex organizational systems is discussed on the example of sustainable development of the region. The factors affecting sustainable development are taken from the materials of the World Conference on Environment and Development of the United Nations. A huge number of factors include the criteria and indicators by which it is possible to evaluate the level of development of a particular geographical region, to make a forecast of its future condition, and to draw a conclusion about the sustainability of this condition. To solve the problem of selecting, out of the variety of factors, the most important (influential) factors affecting the main goal, the following methods are proposed: perceptive-cognitive modeling, statistical analysis, fuzzy sets and graph theory.

2 Structuring the Goal of the System

The goals and factors affecting the complex system are structured, which means that they comprise sub-goals, each of which can have its own sub-goals, and so on. As a result, at the end we will have the list of atomic goals. The number of sub-goals might be too many. We should also take into consideration that particular elements of the atomic goals do not affect the main goal in the same way.

Selection of optimal set of the system goals, as well as the information technology for achieving the system functioning comprises several stages:

Structuring the goal of the system means that the main or global goal of the system functioning is identified, which is assigned a zero level, then it is decomposed into sub-goals.

In order to structure the goal, that is, to create a structured information model, we will identify the main goal (which is a global goal)—sustainable development of the region, which is expressed as— C_0 and to which we assign a zero level. Then, i.e. at the first level, it is decomposed into sub-goals C_1 , C_2 , C_3 and C_4 , and the second level sub-goals are divided into further sub-goals (C_{11} , C_{12} , ..., C_{21} , ... C_{31} ...). Table 1 shows the values and sub-goals of the factors contributing to the sustainable development of the region.

Table 1 The values and sub-goals of the factors

N	Symbol	Value
1	C_0	Sustainable development of the region
2	C_1	Social factor
3	C_2	Economic indicator
4	C_3	Ecological indicator
5	C_4	Organizational factor
6	C_{11}	Fighting against poverty (%)
7	C_{12}	Demographic dynamics (%)
8	C_{13}	Promotion of education, staff training and public information (%)
9	C_{14}	Protection of population health (%)
10	C_{15}	Supporting sustainable development of the population (%)
11	C_{111}	Population employment growth rate (%)
12	C_{112}	Average salary ratio for women and men (%)
13	C_{113}	Populations below the poverty line (%)
14	C_{114}	The ratio between the income of the rich and the poor people
⋮		
185	C_{411}	Ratification of international agreements on sustainable development

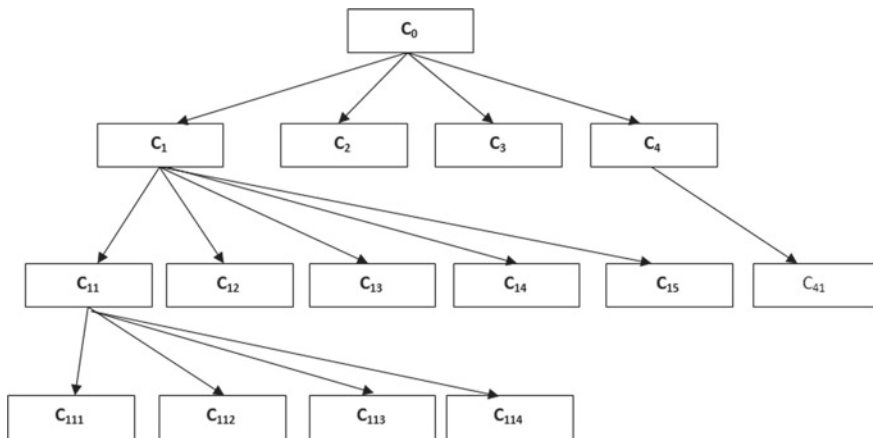


Fig. 1 The goal tree of interdependence of goals and sub-goals

According to these goals and sub-goals, let's build a goal tree that will look as given in Fig. 1.

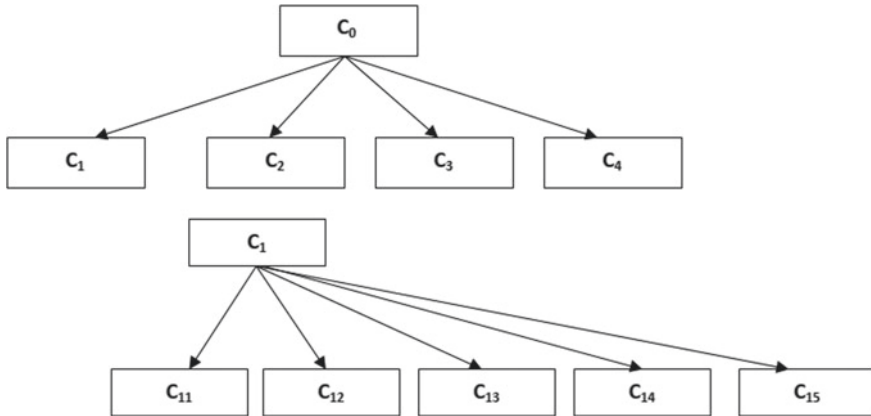


Fig. 2 Zero and first rank (fragment)

Let’s divide two-level fragments from the above, that consist of root tip and incipient tip. We will assign a zero rank to the fragment that contains the root tip of the tree. Lower-level tips of this fragment represent root tips for the first-rank tips. Zero rank is assigned to the fragment from the goal tree (C_0, C_1, C_2, C_3, C_4). The fragments from the goal tree have the first rank: ($C_1, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}$), ($C_2, C_{21}, C_{22}, C_{23}$), ($C_3, C_{31}, C_{32}, C_{33}, C_{34}, C_{35}$), etc. (Fig. 2).

3 Assigning Weight to the Goal of the System

Suppose the multitude of local goals defined at the first stage equals to N , the number of achievement levels of each goal— k , hence, the number of possible solutions, called multipurpose alternative, equals to N^k .

From such a multitude of goals, it is almost impossible to make an optimal choice. Therefore, in order to calculate the effectiveness of achieving the goal (i.e. how effectively the main goal is achieved), we are ranking local goals and select the most important goal from the obtained subgoals; this way we also reduce the number of goal levels.

For ranking goals, each goal is evaluated by numeric value—i.e. by their “weights”, which are assigned by an expert or a group of experts. Such assessment is subjective. The goal is described verbally, which can also include a numeric indicator. This method is called the hierarchy analysis method.

For evaluating the interaction of goals (to introduce weights of tips on the tree), in order to determine how important the goal is, together with the expert we will introduce evaluation scores [1–3]. The interaction power of C_i and C_j goals shall be evaluated verbally (linguistically) and expressed quantitatively in the interval [1–10] (Table 2).

Table 2 The interaction power of C_i and C_j goals

Linguistic meaning	Numeric (points) meaning
C_i and C_j purposes in the same meaning	1
When C_i is weakly dependent on the value of C_j	3
When C_i is strongly dependent on the value of C_j	5
When C_i is very strongly dependent on the value of C_j	7
Absolutely dependent on the value of C_i to C_j	9
Assessment is situated between two linguistic assessments	2, 4, 6, 8

For each fragment of the goal tree (Table 3), we will create a square matrix zero $R = ||r_{ij}||$. The columns of the matrix correspond to the tree tips. In the box of the upper left column, the total weight of the root tip is given (C_0 for global goals weight $W_0 = 1$). At the intersection of C_i line and C_j intersection column r_{ij} value is indicated, this value is equal to 1 if $C_i = C_j$. If C_i is more important than C_j then b_{ij} is indicated, otherwise, if C_i is less important than C_j and $1/b_{ij}$ is indicated. Table 3, the degree of interaction of goals is filled in based on the expert evaluation by using Table 2. The square matrix for the tree fragment is shown in Table 3.

If the rows (columns) of the matrix correspond to the target C_1, \dots, C_p , rated by weights W_1, \dots, W_p . The root tip has a weight W_0 , . then the condition is true [1]:

$$W_q = \sum_{i=1}^p W_i \tag{1}$$

W_1 weights represent the solutions to the following equation systems [1]:

$$\begin{aligned}
 w_1 &= 1/P_1 \sum_{J=1}^P r_{1j} W_j \\
 &\dots \\
 w_{p-1} &= 1/p \sum_{J=1}^P r_{p-1,j} W_j \\
 w_{p-1} &= 1/p \sum_{J=1}^P r_{p-1,j} W_j.
 \end{aligned}
 \tag{2}$$

In the discussed example, the systems of equations of the corresponding 0 fragment are expressed in the following way (3):

$$\begin{aligned}
 w_1 &= 1/4(w_1 + 3w_2 + 3w_3 + 3w_4) \\
 w_2 &= 1/4(1/3w_1 + w_2 + 3w_3 + 3w_4), \\
 w_3 &= 1/4(3w_1 + 1/3w_2 + 3w_3 + 3w_4) \\
 w_4 &= 1 - (w_1 + w_2 + w_3 + w_4)
 \end{aligned}
 \tag{3}$$

From the solution of the system of these equations it will obtain C_1, C_2, C_3, C_4 target weights for goals.

Systems of such equations are drawn for other ranks too. The solution of the system of equations will be obtained $C_{11}, C_{12}, C_{13}, C_{14}, C_{15}...$ Target weights for goals.

4 Minimization of the Local Goals of the System

At the next step it is necessary to carry out numeric evaluations and ranking of the most important goals and factors to select the most effective goal and factor. The most important goals are selected from the goals that have been selected at the first stage by deleting relatively insignificant goals [2].

While minimizing local goals, several conditions shall be simultaneously fulfilled. The following should be taken into consideration:

- Interaction of local goals reflected through the Matrix—Cognitive Map.
- Overall degree of deleted goals having a numeric value and defined by a cognitive map shall not exceed marginal value.
- The number of deleted goals shall be maximum.

To draw a cognitive map of the interdependence of local goals of the factors affecting the sustainable development, it should be taken into account that the columns and rows in the table correspond to local goals, on the right side of the Table C_i goals are included, to the right— W_i weights. At the intersection of columns and rows, an expert evaluation $+\alpha_{ij}$ is written if C_i reinforces achievement of the goal C_j , and $-\alpha_{ij}$ if C_i weakens achievement of the goal C_j , where, $0 \leq \alpha_{ij} \leq 1$.

α_{ij} evaluation may not match the values on the scale and may be in the interval between the values. If the goal does not affect another goal, then $\alpha_{ij} = 0$ and if there is no connection between goals or if the connection is unclear then (C_i, C_j) intersection remains empty.

The numeric values of the interdependence of the goals affecting the factors for sustainable development are given in Table 4.

Given this, the cognitive map will look as it is in Table 5. Fragment of the Cognitive Map

Fragment of Cognitive Map

In order to determine the interdependence of the goals on the cognitive map, we introduce numeric indicators—the degree of achievement of global (C_0) and local goals (C_j), they are calculated by formulas [1]:

Table 6 Table of interactions between goals (fragment)

c_j	c_1	c_2	c_3	c_4	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	
$J(C_j)$	0.0226	0.0201	0.02208	0.011	0.022	0.029	0.0017	0.002	0.005	
	-	-	-	-	-	-	-	-	-	
⋮										
c_j	c_{42}	c_{43}	c_{44}	c_{45}	c_{46}	c_{47}	c_{48}	c_{49}	c_{410}	c_{411}
$J(C_j)$	0.06	0.0201	0.018	0.0031	0.0142	0.0129	0.1117	0.0242	0.025	0.123
	-	-	b_1	-	-	-	-	-	-	b_2

$$\begin{aligned}
 J(C_0) &= \sum_{j=1}^N \sum_{i=1}^N \alpha_{ij} \cdot W_i \\
 J(C_j) &= \frac{\sum_{j=1}^N \alpha_{ij} \cdot W_i}{J(C_0)} = \frac{\sum_{j=1}^N \alpha_{ij} W_i}{\sum_{j=1}^N \sum_{i=1}^N \alpha_{ij} \cdot W_i}
 \end{aligned}
 \tag{4}$$

Which, for the zero-ranking goal of the tree fragment that we have discussed, will look as follows:

$$\begin{aligned}
 (C_0) &= \sum_{j=1}^N (\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) \cdot W_i \\
 &= (\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) \cdot W_1 + (\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) \cdot W_2 \\
 &+ (\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) \cdot W_3 + (\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) \cdot W_4 = 3, 5101
 \end{aligned}
 \tag{5}$$

After performing the above calculations $J(C_0) = 3, 501$, and other values of $J(C_j)$ will obtain the value (Table 6).

We will denote multitude of all local goals as C , multitude of deleted sub-goals— C^* , and their power— $|C^*|$.

The extent of achieved sub-multitude of goals, depending on their interaction is expressed by the formula [1]; $J(C^*) = J(c_{ji} + \dots + J(c_{jk}))$ (6) $J(C^*)$ the maximum permissible value shall be denoted as Δ . And in this case it is equal to 0.2101

Let’s formulate the minimization task: we have to find $C^* C$, so that to accomplish the following conditions simultaneously

$$\begin{aligned}
 J(C^*) &\leq \Delta \\
 |C^*| &= \max
 \end{aligned}
 \tag{6}$$

Given the above, the task of minimizing local goals can be calculated by the following algorithm:

$J(C^*)$ the maximum permissible value shall mark Δ . And it is equal to the discussed case 0.2101

To formulate the minimization task we have to find the following C^*C , So that the following conditions are fulfilled simultaneously

1 In C multitude we will choose such goal C_{ij} that has a minimum degree of achievement ($J(c_{ij}) = \min$). If such goals are more than one, we have to choose any of them. We shall include the selected goal in C^* multitude and increase its degree of achievement.

$$J(C^*) = J(C^*) + J(c_{j1})$$

2 We shall check $J(C^*) \leq \Delta$ condition, if it is accomplished, then we shall delete C_{j1} from C and go back to the first step. If the above condition is not accomplished for any of the goals, i.e. C^* goal cannot be joined to any other goal, it means that the algorithm works.

If for the example we discussed $\Delta = 0, 29$., then the outcome of minimization of local goals is the multitude $E = b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8$, where

b_1 —denotes presence of the national strategy for sustainable development (Yes/No)

b_2 —is international agreements for sustainable development.

b_3 —Share of per capita on the national income (%).

b_4 —Population density (%).

b_5 —Increasing the birth rate (%).

b_6 —Entrepreneurial activity (%).

b_7 —Changing the nature of the request.

b_8 —updated share of national product per capita (%)

b_9 —Export share of national GDP (%).

b_{10} —Share of investment(%).

b_{11} —Populations below the poverty line in drought-prone areas (%)

b_{12} —management of eco-systems, combating desertification and drought.

b_{13} —Population growth rate.

b_{14} —Costs to rebuild the ecosystem (%).

Using the methodological basis we have discussed below, we have selected relatively high-level factors of sustainable development whose degree (contribution) to sustainable development will be significant.

5 Conclusion

The given paper deals with the methods of perceptive-cognitive modeling, statistical analysis, fuzzy set theory and graph theory to manage complex organizational systems and make decisions, on the example of managing the sustainable development of the region. To solve the given problem, the main goal was identified, a tree of goals

and sub-goals was built; a structured information model was created. To make the optimal choice, a mathematical model was developed and the weight of each goal was determined; the goals of the system were minimized and from the multitude of goals, only those goals were selected whose impact on the main goal is more important. This, in turn, is crucial to ensure sustainable development of the region, to manage the process and make decisions.

References

1. Uditsky, S.A., Vladislavlev, P.N.: The Basics of Predesign Analysis of the Organizational Systems . M. "Finances and Statistics" (2005)
2. Gasitashvili, Z., Rodonaia, Z.I., Kiknadze, I.M.: Building of research model for region stable development. In: XII ALL—Russian Meeting on Management Problems of VSPU-2014 June 16–19 (2014)
3. Vladislavlev, P.N.: Selection of the optimal scenario of behavior of the organizational systems. In: "Management of the Major Projects ". Collection of papers. Ed by D.A. Novikov PMI, M. (2005)
4. Gubko, M.V., Novmkov, D.A.: Game Theory in the Management of the Organizational Systems(2002)

On a Generalized Integro-Differential Spatial Model of Economic Growth



Herb Kunze, Davide La Torre, and Simone Marsiglio

Abstract We analyze a spatial economic growth model in which production in each location is determined by the amount of output produced in other locations within an industrial cluster as well. Therefore, the evolution of capital gives rise to an integro-differential extension of the basic spatial economic growth model. We analyze the model both in a purely dynamic setting and in an optimal control framework, proposing a numerical algorithm to solve the model under the latter scenario. Different from previous studies, our algorithm allows to solve the model even in a setting in which the objective function is nonlinear, permitting thus to analyze the spatial features of the model even in its traditional formulation from economic growth theory.

Keywords Economic growth · Spatial Solow's model · Ramsey model · Cobb-Douglas production function

1 Introduction

Economic growth models have been recently extended to a spatial dimension in order to characterize how different locations within the whole economy interact with one another through the trade channel [1, 4–6, 9]. Most of the papers focus on a Solow-type [14] purely dynamic setting in which agents' behavior is exogenously given, while more limited in number are those analyzing a Ramsey-type [13] setting

H. Kunze (✉)

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

D. La Torre

SKEMA Business School - Universite' de la Cote-d'Azur, Sophia Antipolis, France
e-mail: davide.latorre@skema.edu

S. Marsiglio

University of Pisa, Pisa, Italy
e-mail: simone.marsiglio@unipi.it

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_62

in which agents optimally determine their actions. In particular, the presence of agents' optimization precludes to derive analytical results unless in specific model's formulation in which either the objective function or the dynamic constraint is linear [3], and even the numerical analysis is not simple to manage since the problem turns out to be ill-posed unless specific restrictive assumptions are made [2]. Our contribution to this literature is twofold: we present an extended spatial economic growth model to describe the production process within an industrial cluster in which the output produced in a given location strictly depends on the output produced in other locations as well; we present a numerical algorithm which is general enough to determine the spatio-temporal evolution of the the main variables without imposing specific restrictions on the shape of the objective function and the dynamic constraint.

The paper proceeds as follows. Section 2 discusses the model in a purely dynamic setting characterizing both analytically and numerically some interesting results. Section 3 introduces agents' optimization in the model analyzing numerically the spatio-temporal behavior of the control and state variables. Section 6 as usual concludes.

2 The Solow-Type Model

Given a compact set $\Omega \subset R^n$, consider the following integro-differential extension of the spatial Solow model with Neumann boundary conditions on Ω :

$$\begin{cases} \frac{\partial K(x,t)}{\partial t} = \nabla(d_k(x)\nabla K(x,t)) + \int_{\Omega} \phi(x,y)K(y,t)^{\alpha}dy - \delta_K K(x,t), & (x,t) \in \Omega \times (0, +\infty) \\ d_k(x) \frac{\partial K}{\partial n}(x) = 0, & x \in \partial\Omega \\ K(x,0) = K_0(x). & x \in \Omega \end{cases}$$

where $K(x,t)$ is the capital stock at time t in location x , $K_0(x)$ is the initial distribution of capital, $\alpha \in (0, 1]$ is the capital share, and $\phi(x,y)$ is a positive kernel. This model represents an extension of the following spatial Solow model:

$$\begin{cases} \frac{\partial K(x,t)}{\partial t} = \nabla(d_k(x)\nabla K(x,t)) + A(x)K(x,t)^{\alpha} - \delta_K K(x,t), & (x,t) \in \Omega \times (0, +\infty) \\ d_k(x) \frac{\partial K}{\partial n}(x) = 0, & x \in \partial\Omega \\ K(x,0) = K_0(x). & x \in \Omega \end{cases}$$

which has been studied, for instance, in [2] and that can be obtained as a particular case of our formulation by taking $\phi(x,y)$ equal to the modified Dirac $A(y)\delta_x(y)$. For the sake of simplicity the population size and the saving rate have been normalized to unity. In such a spatial formulation, following [9] and [11], we interpret any location x as a single local entity within the entire economy, in order to allow for the existence of heterogeneities between different local entities. More specifically, the entire economy may represent an industrial clusters in which different local entities

are all related from buyer-supplier relationships ([15]), meaning that the amount of output produced in a given location strictly depends on the amount of output produced in other locations as well. The importance of agglomeration and industrial clusters for economic activities in general and economic growth in particular have been extensively discussed, both from theoretical and empirical points of view ([8]; [12]). However, to the best of our knowledge, the peculiarities of industrial clusters have never been specifically accounted for in spatial growth models, which give rise to the natural framework to discuss them. Our goal in this paper is therefore to move a first step in this direction by allowing the spatial domain to represent an industrial cluster in which all cluster agents interact. In order to do so, we need to distinguish the output produced at given location and the total output available for sale and consumption at that same location, which results from the aggregation of the output produced within the entire cluster. The output produced at location x is determined by a Cobb-Douglas production function, $Q(x, t) = K(x, t)^\alpha$, while the total output at location x is obtained by averaging the output produced at each $y \in \Omega$ through a weight $\phi(x, y)$ as follows: $Y(x, t) = \int_\Omega \phi(x, y)Q(y, t)dy$, where the kernel $\phi(x, y)$ measures the degree of interrelations between different localities within the cluster. In our spatial setting, the peculiarities of the cluster are captured by spatial diffusion, which accounts for the trade relations between localities, and the integral term, which accounts for the production relations between localities. These two elements play a diametrically different role in determining the spatio-temporal dynamics: while spatial diffusion acts as a convergence mechanism which tends to smooth out all spatial heterogeneities over time, the integral term acts as a divergence mechanism which tends to reinforce the presence of spatial heterogeneities ([9]).

In this well known that a closed-form solution to this model exists in 1-d when $d_k(x) = d_k$, ϕ is the Dirac concentrated at x , and $\alpha = 1$.

Definition 1 An equilibrium of the above system is a function $\bar{K}(x)$ that is a solution to the following system:

$$\begin{cases} \nabla (d_k(x)\nabla K(x)) + \int_\Omega \phi(x, y)K(y)^\alpha dy - \delta_K K(x) = 0, & x \in \Omega \\ d_k(x) \frac{\partial K}{\partial n}(x) = 0. & x \in \partial\Omega \end{cases}$$

Before proceeding, let us notice that there are cases in which it can happen that $\int_\Omega \phi(x, y)dy$ is a positive constant, not depending on x . This is the case, for instance, when $\phi(x, y) = \delta_x(y)$ or $\phi(x, y) = \frac{1}{\mu(\Omega)}$. The following result states the existence of two equilibria (homogeneous) for the above system.

Proposition 1 Let $\alpha \in (0, 1)$ and suppose that $\int_\Omega \phi(x, y)dy = 1$ for any $x \in \Omega$. Then $K_1(x) = 0$ and $K_2(x) = \delta_K^{-\frac{1}{1-\alpha}}$ are the only two homogeneous equilibria of the above PDE. If $\alpha = 1$ then the above equation has only the equilibrium $K_1(x) = 0$.

Proof It is easy to show that K_1 and K_2 are two homogeneous solutions to the above equation. It is also trivial to prove that there are no other homogeneous equilibria of the above system.

In a more general context, we can provide the following upper bound of the solution. Let us introduce the total amount of capital over Ω as follows:

$$\xi(t) = \int_{\Omega} K(x, t) dx \tag{1}$$

The following result provides an upper bound for ξ , suggesting that the total amount of capital within the whole economy cannot grow indefinitely and its upper bound depends on the main model’s parameters.

Proposition 2 *Suppose that ϕ is bounded over Ω by a constant $\theta > 0$ and let $\mu(\Omega)$ be the measure of the compact set Ω . Then we have the following upper bound estimate of ξ :*

$$\xi(t) \leq \left[\frac{\theta \mu(\Omega)^{2-\alpha} (e^{(1-\alpha)\delta_K t} - 1) + \delta_K \phi(0)^{1-\alpha}}{\delta_K e^{(1-\alpha)\delta_K t}} \right]^{\frac{1}{1-\alpha}}$$

Proof Computing we have

$$\begin{aligned} \xi'(t) &= \int_{\Omega} \frac{\partial K(x, t)}{\partial t} dx \\ &= \int_{\Omega} \nabla (d_k(x) \nabla K(x, t)) + \int_{\Omega} \int_{\Omega} \phi(x, y) K(y, t)^{\alpha} dy dx - \delta_K K(x, t) dx \\ &= \int_{\Omega} \nabla (d_k(x) \nabla K(x, t)) + \int_{\Omega} K(y, t)^{\alpha} \left(\int_{\Omega} \phi(x, y) dx \right) dy - \int_{\Omega} \delta_K K(x, t) dx \end{aligned}$$

and by using Jensen’s inequality we get:

$$\begin{aligned} \xi'(t) &= \mu(\Omega)^2 \theta \left[\frac{1}{\mu(\Omega)} \int_{\Omega} K(y, t)^{\alpha} dy \right] - \delta_K K(x, t) dx \\ &\leq \mu(\Omega) \theta \left(\frac{1}{\mu(\Omega)} \int_{\Omega} K(y, t) dy \right)^{\alpha} - \delta_K \int_{\Omega} K(x, t) dx \\ &= \theta \mu(\Omega)^{2-\alpha} \xi(t)^{\alpha} - \delta_K \xi(t) \end{aligned}$$

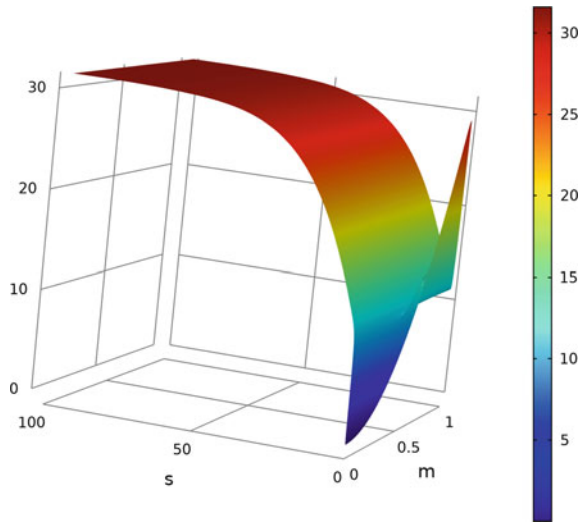
which implies that ϕ satisfies the following inequality

$$\xi'(t) + \delta_K \xi(t) \leq \theta \mu(\Omega)^{2-\alpha} \xi(t)^{\alpha} \tag{2}$$

that is

$$\xi(t) \leq \left[\frac{\theta \mu(\Omega)^{2-\alpha} (e^{(1-\alpha)\delta_K t} - 1) + \phi(0)^{1-\alpha}}{e^{(1-\alpha)\delta_K t}} \right]^{\frac{1}{1-\alpha}}$$

Fig. 1 Evolution of physical capital



We now present some simple numerical simulations to exemplify the spatio-temporal evolution of capital. We consider two different model’s parametrizations in which the initial distribution of capital is heterogenous over space.

Example 1 We choose $\phi(x, y) = 1$, $d_k(x) = x$, $K_0(x) = 30x^2$, $\delta_K = 0.1$, $\alpha = \frac{1}{3}$. The following Fig. 1 shows the evolution of physical capital over space and time.

Example 2 We choose $\phi(x)$ equal to the Dirac at x , $d_k(x) = x$, $K_0(x) = 30x^2$, $\delta_K = 0.1$, $\alpha = \frac{1}{3}$. The following Fig. 2 shows the evolution of physical capital over space and time.

In both cases, despite the heterogenous initial distribution of capital, the long-run evolution of the capital shows the convergence to a non-trivial and homogeneous equilibrium, suggesting that the convergence effects associated with diffusion more than offset the divergence effects associated with the integral term ([9]). The following example shows that non-homogeneous equilibria can exist.

Example 3 Let $\Omega = [-\frac{\pi}{2}, \frac{\pi}{2}]$, $d_k(x) = 1$, $\delta_K = 1$, and $\alpha = 1$. Let

$$\phi(x, y) = \frac{1}{\pi} \left[\frac{2 + 2 \sin x}{2 + \sin y} \right] \geq 0$$

It is easy to prove that the function $\bar{K}(x) = 2 + \sin x$ is a steady-state solution (see Fig. 3).

Fig. 2 Evolution of physical capital

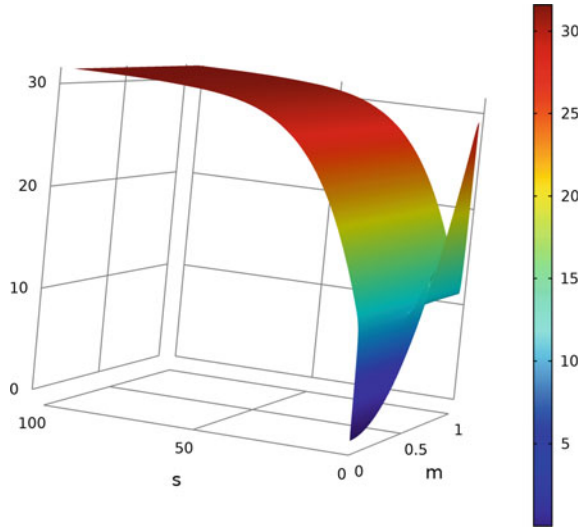
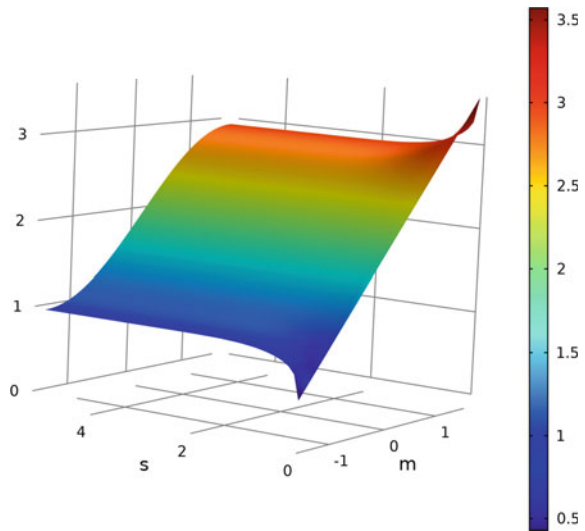


Fig. 3 Evolution of physical capital



3 The Ramsey-Type Model

We now introduce agents optimization by assuming that the whole economy is ruled by a social planner which determines the level of consumption $C(x, t)$ in each location in order to maximize social welfare (defined as the discounted sum of utilities) adjusted for sustainability considerations. The model can thus be written as the following optimal control problem:

$$\max J(C, K) = \int_0^T \int_{\Omega} U(C(x, t))e^{-\rho t} dxdt + \Theta \int_{\Omega} K(x, T)dx$$

Subject to

$$\begin{cases} \frac{\partial K(x,t)}{\partial t} = \nabla(d_k(x)\nabla K(x, t)) + \int_{\Omega} \phi(x, y)K(y, t)^{\alpha}dy - C(x, t) - \delta_K K(x, t), & (x, t) \in \Omega \times (0, +\infty) \\ d(x) \frac{\partial K}{\partial n}(x) = d(x) \frac{\partial K}{\partial n}(x) = 0, & x \in \partial\Omega \\ K(x, 0) = K_0(x). & x \in \Omega \end{cases}$$

where U is the utility function. As usual, U is supposed to be increasing and concave. The objective function $J(C, K)$ is the sum of two terms with $\Theta > 0$ measuring the importance of the second term relative to the first. While the first term describes the level of welfare in the whole economy, the second accounts for the sustainability concerns associated with the capital level remaining at the end of the planning horizon T for later generations [7, 10]. To determine an optimal policy result, let us define the current Hamiltonian function as

$$H(C, K, \lambda) = U(C) + \lambda_K(\nabla(d_k(x)\nabla K(x, t)) + \nabla(d_k(x)\nabla K(x, t)) + \int_{\Omega} \phi(x, y)K(y, t)^{\alpha}dy - \delta_K K(x, t) - C(x, t))$$

The following proposition provides the optimality conditions for an optimal solution of the problem above.

Proposition 3 *A pair (\tilde{C}, \tilde{K}) solves the above optimal control model if and only if it is solution to the following Hamiltonian system:*

$$\begin{cases} \frac{\partial K(x,t)}{\partial t} = \nabla(d_k(x)\nabla K(x, t)) + \int_{\Omega} \phi(x, y)K(y, t)^{\alpha}dy - \delta_K K(x, t) - C(x, t), & (x, t) \in \Omega \times (0, T) \\ \frac{\partial \lambda(x,t)}{\partial t} = \rho\lambda - \nabla(d_k(x)\nabla \lambda(x, t)) - \lambda\alpha \int_{\Omega} \phi(x, x')K^{\alpha-1}(x', t) - \delta_K \lambda, & (x, t) \in \Omega \times (0, T) \\ C(x, t) = \lambda(x, t)^{-\frac{1}{\theta}} & (x, t) \in \Omega \times (0, T) \\ d(x) \frac{\partial K}{\partial n}(x) = 0, & x \in \partial\Omega \\ d(x) \frac{\partial \lambda}{\partial n}(x) = 0, & x \in \partial\Omega \\ \lambda(x, T) = \Phi & x \in \Omega \\ K(x, 0) = K_0(x) & x \in \Omega \end{cases}$$

4 The Maximization Algorithm

Since analyzing explicitly the Hamiltonian system above is generally not possible (unless we introduce restrictive assumptions), we now proceed with numerical simulations to illustrate the optimal behavior of capital and consumption.

If we use the dynamic constraint and plug it into the objective function we obtain:

$$J(K) = \int_0^T \int_{\Omega} U \left(-\frac{\partial K(x,t)}{\partial t} + \nabla(d_k(x)\nabla K(x,t)) + \int_{\Omega} \phi(x,y)K(y,t)^\alpha dy - \delta_K K(x,t) \right) e^{-\rho t} dx dt + \Theta \int_{\Omega} K(x,T) dx$$

Subject to

$$\begin{cases} -\frac{\partial K(x,t)}{\partial t} + \nabla(d_k(x)\nabla K(x,t)) + \int_{\Omega} \phi(x,y)K(y,t)^\alpha dy - \delta_K K(x,t) \geq 0, & (x,t) \in \Omega \times (0,T) \\ d(x) \frac{\partial K}{\partial n}(x) = 0, & x \in \partial\Omega \\ K(x,0) = K_0(x). & x \in \Omega \end{cases}$$

The direction derivative of J along any feasible h is given by

$$\begin{aligned} J'(K; h) &= \lim_{\delta \rightarrow 0} \frac{J(K + \delta h) - J(K)}{\delta} \\ &= \int_0^T \int_{\Omega} U' \left(-\frac{\partial K(x,t)}{\partial t} + \nabla(d_k(x)\nabla K(x,t)) + \int_{\Omega} \phi(x,x')K^\alpha(x',t) dx' - \delta_K K(x,t) \right) \\ &\quad * \left(-\frac{\partial h(x,t)}{\partial t} + \nabla(d_k(x)\nabla h(x,t)) + \int_{\Omega} \phi(x,x')h^\alpha(x',t) dx' - \delta_K h(x,t) \right) e^{-\rho t} dx dt \\ &\quad + \Theta \int_{x_a}^{x_b} h(x,T) dx \end{aligned}$$

We propose an algorithm to determine an approximation of the optimal solution. At each step this algorithm determines the direction of growth h using the above calculated directional derivative $J(K; h)$.

(1) Given the value of the state variable $K_n(x, t)$, solve the following problem

$$\begin{cases} -\frac{\partial h(x,t)}{\partial t} + \nabla(d_k(x)\nabla h(x,t)) + \int_{\Omega} \phi(x,x')h^\alpha(x',t) dx' - \delta_K h(x,t) = \\ \left[U' \left(-\frac{\partial K_n(x,t)}{\partial t} + \nabla(d_k(x)\nabla K_n(x,t)) + \int_{\Omega} \phi(x,x')K_n^\alpha(x',t) dx' - \delta_K K_n(x,t) \right) \right]^{-1} \\ \times (-\Theta \frac{\partial h}{\partial t} + 1) e^{\rho t}, & (x,t) \in \Omega \times (0,T) \\ d(x) \frac{\partial h}{\partial n}(x) = 0, & x \in \partial\Omega \\ h(x,0) = 0. & x \in \Omega \end{cases}$$

(2) Determine $\delta > 0$ that corresponds to the maximum increment of J along the direction h

(3) Update $K_{n+1} = K_n + \delta h$

(4) If $|J(K_{n+1}) - J(K_n)| < \epsilon$ then stop otherwise go to point (1).

The following result shows that J is increasing along the sequence generated by the above algorithm. The implementation of the above algorithm generates a sequence of functions K_n along which the objective function is increasing.

Proposition 4 *If δ is small then $J(K_{n+1}) \geq J(K_n), \forall n \geq 0$.*

Proof Computing we have:

$$\begin{aligned}
 J(K_{n+1}) - J(K_n) &= \delta J'(K_n; h) + o(\delta) \\
 &= \delta \left(\int_0^T \int_{\Omega} \left(-\Theta \frac{\partial h}{\partial t} + 1 \right) dxdt + \Theta \int_{\Omega} h(x, T) dx \right) + o(\delta) \\
 &= \delta \left(-\Theta \int_{\Omega} h(x, T) - h(x, 0) dx + T\mu(\Omega) + \Theta \int_{\Omega} h(x, T) dx \right) + o(\delta) \\
 &= \delta \left(T\mu(\Omega) + \frac{o(\delta)}{\delta} \right) \geq 0
 \end{aligned}$$

and this last passage relies on the boundary condition $h(x, 0) = 0$.

5 Numerical Examples

We now run a numerical simulation by relying of the above described algorithm assuming that U takes the form:

$$U(C) = (1 + C)^\theta - 1.$$

This function satisfies the classical hypotheses which define an utility function (i.e. is increasing and concave).

Example 4 In this example we suppose $K_0(x) = 1 + x, d_K = 1 - 0.5x^2, \Theta = 0, \alpha = 1, \theta = \frac{2}{3}$, and $\phi(x, y) = \delta_x(y)$. The following Fig. 4 shows the long-run behavior of K . The values of the objective function after three iterations are: $J_0 = 0.579453721074241, J_1 = 0.6059822543917376, J_2 = 0.6287663921318654, J_3 = 0.6534865860743782$.

Example 5 In this example we suppose $K_0(x) = 1 + x, d_K = 1 - 0.5x^2, \Theta = 0.1, \alpha = 1, \theta = \frac{2}{3}$, and $\phi(x, y) = \delta_x(y)$. The following Fig. 5 shows the long-run behavior of K . The values of the objective functions after three iterations are given by: $J_0 = 0.5944537210742411, J_1 = 0.6200510708312351, J_2 = 0.6404132122774402$, and $J_3 = 0.6570096551286734$.

Example 6 In this example we suppose $K_0(x) = 1 + x, d_K = 1 - 0.5x^2, \Theta = 0.1, \alpha = \theta = \frac{2}{3}$, and $\phi(x, y) = 2\delta_x(y)$. The following Fig. 6 shows the long-run behavior of K . The values of the objective functions after three iterations are given by: $J_0 = 1.3119041162656826, J_1 = 1.3265311280015393$.

Fig. 4 Evolution of physical capital

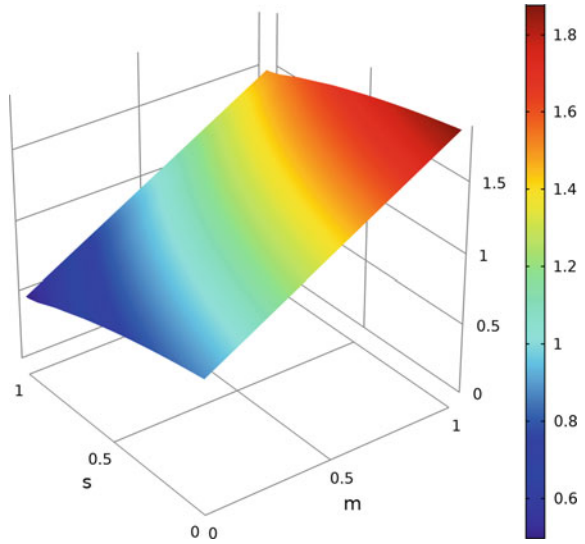
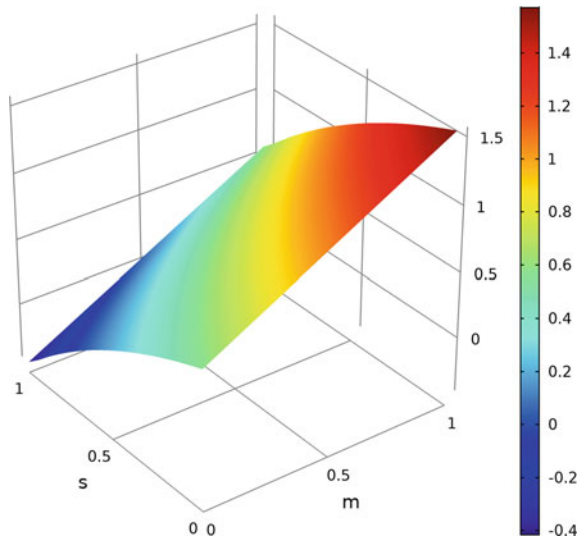
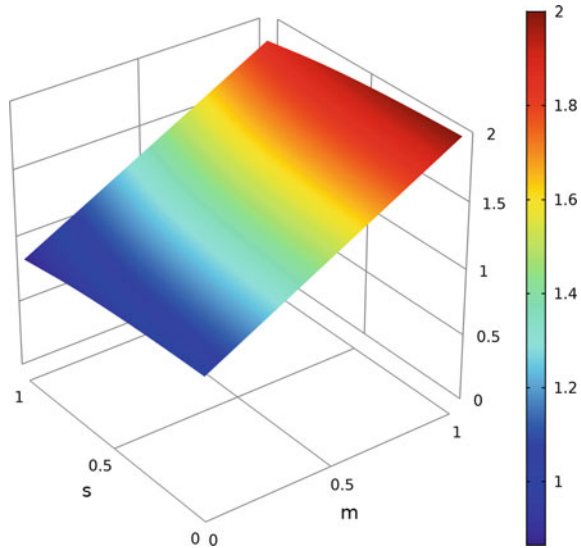


Fig. 5 Evolution of physical capital



The above figures illustrate that our simple algorithm is capable to handle an optimal control problem on partial differential equations without requiring specific restrictive assumptions on the functional forms of the utility and production functions.

Fig. 6 Evolution of physical capital



6 Conclusion

The importance of agglomeration and industrial clusters for economic growth has been extensively documented. However, the peculiarities of clusters have never been specifically accounted for in spatial growth models, despite they represent the natural framework to discuss them. In order to move a first step in this direction, in this paper we analyze a spatial economic growth model within an industrial cluster. The total output in each location is determined by the amount of output produced in other locations within the cluster, and so production activities within the cluster are all interrelated. This implies that the evolution of capital gives rise to a spatial integro-differential equation, which represents a generalization of the basic spatial economic growth model. We analyze the model both in a purely dynamic setting and in an optimal control framework, proposing a numerical algorithm to solve the model under the latter scenario. Different from previous studies, our algorithm allows us to solve the model even in a setting in which the objective function is nonlinear, permitting thus to analyze the spatial features of the model even in its traditional formulation from economic growth theory.

References

1. Boucekkine, R., Camacho, C., Zou, B.: Bridging the gap between growth theory and economic geography: the spatial Ramsey model. *Macroecon. Dyn.* **13**, 20–45 (2009)
2. Boucekkine, R., Camacho, C., Fabbri, G.: On the optimal control of some parabolic differential equations arising in economics. *Serdica Math. J.* **39**, 331–354 (2013a)

3. Boucekkiné, R., Camacho, C., Fabbri, G.: Spatial dynamics and convergence: the spatial AK model. *J. Econ. Theor.* **148**, 2719–2736 (2013b)
4. Brito, P.: The dynamics of growth and distribution in a spatially heterogeneous world. Technical University of Lisbon, UECE-ISEG (2004)
5. Camacho, C., Zou, B.: The spatial Solow model. *Econ. Bull.* **18**, 1–11 (2004)
6. Camacho, C., Zou, B., Briani, M.: On the dynamics of capital accumulation across space. *Eur. J. Oper. Res.* **186**(2), 451–465 (2008)
7. Colapinto, C., Liuzzi, D., Marsiglio, S.: Sustainability and intertemporal equity: a multicriteria approach. *Ann. Oper. Res.* **251**, 271–284 (2017)
8. Krugman, P.: Increasing returns and economic geography. *J. Polit. Econ.* **99**, 483–499 (1991)
9. La Torre, D., Liuzzi, D., Marsiglio, S.: Pollution diffusion and abatement activities across space and over time. *Math. Soc. Sci.* **78**, 48–63 (2015)
10. La Torre, D., Liuzzi, D., Marsiglio, S.: Pollution control under uncertainty and sustainability concern. *Environ. Resour. Econ.* **67**, 885–903 (2017)
11. La Torre, D., Liuzzi, D., Marsiglio, S.: Population and geography do matter for sustainable development. *Environ. Dev. Econ.* **24**, 201–223 (2019)
12. Quah, D.T.: Regional convergence clusters across Europe. *Eur. Econ. Rev.* **40**, 951–958 (1996)
13. Ramsey, F.: A mathematical theory of saving. *Econ. J.* **38**, 543–559 (1928)
14. Solow, R.M.: A contribution to the theory of economic growth. *Q. J. Econ.* **70**, 65–94 (1956)
15. Wolfe, D.A., Gertler, M.S.: Clusters from the inside and out: local dynamics and global linkages. *Urban Stud.* **41**, 1071–1093 (2004)

Utilizing Bidirectional Encoder Representations from Transformers for Answer Selection



Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang

Abstract Pre-training a transformer-based model for the language modeling task in a large dataset and then fine-tuning it for downstream tasks has been found very useful in recent years. One major advantage of such pre-trained language models is that they can effectively absorb the context of each word in a sentence. However, for tasks such as the answer selection task, the pre-trained language models have not been extensively used yet. To investigate their effectiveness in such tasks, in this paper, we adopt the pre-trained Bidirectional Encoder Representations from Transformer (BERT) language model and fine-tune it on two Question Answering (QA) datasets and three Community Question Answering (CQA) datasets for the answer selection task. We find that fine-tuning the BERT model for the answer selection task is very effective and observe a maximum improvement of 13.1% in the QA datasets and 18.7% in the CQA datasets compared to the previous state-of-the-art.

Keywords Answer selection · BERT · Transformer · Question answering · Deep learning · Machine learning

1 Introduction

Answer Selection is a fundamental problem in the areas of Information Retrieval and Natural Language Processing (NLP) [32]. In the answer selection task, a question along with a list of candidate answers are given and the objective is to rank these

M. T. R. Laskar (✉)

Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
e-mail: tahmedge@cse.yorku.ca

M. T. R. Laskar · J. Xiangji Huang

School of Information Technology, Information Retrieval & Knowledge Management Research Lab, York University, Toronto, Canada
e-mail: jhuang@yorku.ca

E. Hoque

School of Information Technology, York University, Toronto, Canada
e-mail: enamulh@yorku.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_63

Table 1 An example of the Answer Selection task. A question along with a list of candidate answers are given. The sentence in bold font is the correct answer

Question:

- Who is the winner of the US Open 2019?

List of Candidate Answers:

- Rafael Nadal has won the French Open 2019.
- **Rafael Nadal has won the US Open 2019.**
- Roger Federer has won the Australian Open 2018.

Potential Ranking:

- **Rafael Nadal has won the US Open 2019.**
 - Rafael Nadal has won the French Open 2019.
 - Roger Federer has won the Australian Open 2018.
-

candidate answers based on their relevance with the given question [11] (see Table 1). In such tasks, the relevance between a question and a candidate answer is measured by various sentence similarity modeling techniques [32].

In recent years, various sentence similarity models based on the neural network architecture have been utilized to measure the similarity between the question and the candidate answer [2, 3, 23]. In such neural models, first, the word embedding (GloVe [21] or Word2Vec [19]) representations of the question and the candidate answer are used as input to the model. Then the vector representations of these sentences produced by the neural model are utilized for the similarity calculation [2, 3]. However, such word embeddings can only provide a fixed representation of a word and fail to capture its context. Very recently, pre-trained language models have received a lot of attention as they can provide contextual representations of each word in different sentences [5, 22]. Among the pre-trained language models, fine-tuning the transformer-based [28] BERT model yields state-of-the-art performance across different NLP tasks [5]. However, the fine-tuned BERT model is not deeply investigated for the answer selection task yet [11].

To be noted that, there are some issues to address regarding fine-tuning a pre-trained model in a new dataset. For instance, the BERT model has been pre-trained in two scenarios: a) when casing information was present, and b) when casing information was absent. Since it is not guaranteed that all datasets will have conventional casing information, it is important to build models that are robust in scenarios when casing information is missing [18]. In addition, it has been observed that neural models which are trained in datasets having conventional casing perform very poorly in the test data for tasks such as named entity recognition [1] when the conventional casing is absent [18]. Thus, to address the above issues, in this paper, we fine-tune both the cased and uncased versions of the BERT model for the answer selection task. More concretely, our contributions presented in this paper are the following:

- First, we conduct extensive experiments in five datasets by fine-tuning the BERT model for the answer selection task and observe that the fine-tuned BERT model outperforms all prior work where pre-trained language models were not utilized.
- Second, we show that the cased model of BERT for answer selection is as effective as its uncased counterpart in scenarios when casing information is absent.
- Finally, we conduct ablation study to further investigate the effectiveness of fine-tuning BERT for answer selection. As a secondary contribution, we have made our source codes publicly available here: <https://github.com/tahmedge/BERT-for-Answer-Selection>.

2 Related Work

Earlier, various feature engineering-based approaches have been utilized for the answer selection task [24, 32]. However, the feature engineering-based approaches require lots of handcrafted rules and are often error-prone [3]. Also, the features which are used in one dataset are not robust in other datasets [3].

In recent years, several models based on deep neural network have been applied for the answer selection task and they showed impressive performance without requiring any handcrafted features [2–4, 9, 23, 26]. To be noted that, these deep neural network models for answer selection mostly utilized the Recurrent Neural Network (RNN) architecture. However, very recently, models based on the transformer architecture [28] have outperformed the previously proposed RNN-based models in several NLP tasks [5, 17]. Though these transformer-based models utilized the pre-trained BERT architecture [5], models based on BERT have not been deeply investigated for the answer selection task yet. Moreover, it was found that neural models trained on case sensitive texts performed poorly in scenarios when the conventional casing was missing in the test data [18]. Therefore, to address these issues, we utilize both the cased and uncased versions of the pre-trained BERT model and investigate its generalized effectiveness by conducting extensive experiments in five answer selection datasets.

3 Utilizing BERT for Answer Selection

In this section, we first discuss the transformer encoder [28] which was utilized in BERT [5]. Then we briefly describe how the BERT model was pre-trained, followed by demonstrating our approach of fine-tuning the pre-trained BERT model for the answer selection task.

3.1 *Transformer Encoder*

The transformer model has an encoder which reads the text input and a decoder which produces the predicted output of the input text [28]. The BERT model only utilizes the encoder of transformer [5]. The transformer encoder uses the self-attention mechanism to represent each token in a sentence based on other tokens. This self-attention mechanism works by creating three vectors for each token, which are: a query vector Q , a key vector K , and a value vector V . These three vectors are created by multiplying the embedding vector x_i with three weight matrices (W_Q , W_K , W_V) respectively. If d_k is the dimension of the key and query vectors, then the output Z of self-attention for each word is calculated based on the following:

$$Z = \text{softmax} \left(\frac{Q \times K^T}{\sqrt{d_k}} \right) V \quad (1)$$

Since the transformer encoder uses multi-head attention mechanism to give attention on different positions, the self attention is computed eight times with eight different weight matrices which provide eight Z matrices. Then the eight Z matrices are concatenated into a single matrix which is later multiplied with an additional weight matrix in order to send the resulting matrix to a feed-forward layer [28].

3.2 *Pre-training the BERT Model*

The BERT model adopts the encoder of the transformer architecture [28]. The encoder of BERT was pre-trained for masked language modeling and the next sentence prediction task on the BooksCorpus (800M words) [35] dataset along with the English Wikipedia (2,500M words) [5]. For the masked language modeling task, 15% tokens in each input sequence are replaced with the special [MASK] token. The model then learns to predict the original value of the masked words based on the context provided by the non-masked words in the input sequence. In the next sentence prediction task, the model receives a pair of sentences as input and attempts to predict if the second sentence in the input pair is a subsequent sentence in the original document.

3.3 *Fine-Tuning BERT for Answer Selection*

Let's assume that we have two sentences $X = x_1, x_2, \dots, x_m$ and $Y = y_1, y_2, \dots, y_n$. To input them into the BERT model, they are combined together into a single sequence where a special token [SEP] is added at the end of each sentence. Another special

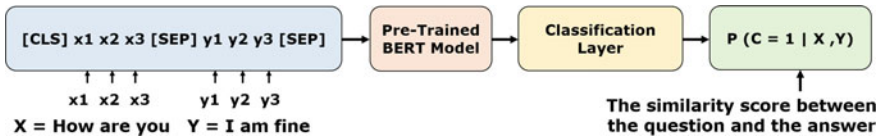


Fig. 1 BERT Fine Tuning: The question X and the candidate answer Y are combined together as input to the pre-trained BERT model for fine-tuning

token $[CLS]$ is added at the beginning of the sequence. The fine-tuning process of the BERT model for the answer selection task is shown in Fig. 1.

In the fine-tuned BERT model, the representation of the first token ($[CLS]$), which is regarded as the aggregate representation of the sequence, is considered as the output of the classification layer. For fine-tuning, parameters are added to the pre-trained BERT model for the additional classification layer W . All the parameters of the pre-trained BERT model along with the additional parameters for the classifier W are fine-tuned jointly to maximize the log-probability of the correct label. The probability of each label $P \in \mathbb{R}^K$ (where K is the total number of classifier labels) is calculated as follows:

$$P = \text{softmax}(CW^T) \tag{2}$$

In the answer selection task, there are two classifier labels (where 1 indicates that the candidate answer is relevant to the question, and 0 indicates the opposite). In the original BERT model [5], sentence pair classification task was done by determining the correct label. But in this paper, we modify the final layer by following the approach of [11] and only consider the predicted score P_{tr} for the similarity label to rank the answers based on their similarities with the question.

$$P_{tr} = P(C = 1|X, Y) \tag{3}$$

4 Experimental Setup

In this section, we present the datasets, the training parameters, and the evaluation metrics used in our experiments. To note that all experiments were run using Nvidia V100 with 4 GPUs.

4.1 Datasets

In our experiments, we used five datasets: two of them were Question Answering (QA) datasets whereas rest were Community Question Answering (CQA) datasets.

Table 2 An overview of the datasets used in our experiments. Here, “#” denotes “number of”

Dataset	# Questions			# Candidate answers		
	Train	Valid	Test	Train	Valid	Test
TREC-QA	1229	82	100	53417	1148	1517
WikiQA	873	126	243	8672	1130	2351
YahooCQA	50112	6289	6283	253440	31680	31680
SemEval-2016CQA	4879	244	327	36198	2440	3270
SemEval-2017CQA	4879	244	293	36198	2440	2930

The overall statistics of the datasets are shown in Table 2. In the following, we give a brief description of each dataset.

TREC-QA: The TREC-QA dataset is created from the QA track (8–13) of Text REtrieval Conference [29].

WikiQA: The WikiQA is an open domain QA dataset [31] in which the answers were collected from the Wikipedia.

YahooCQA: The YahooCQA¹ dataset is a community-based question answering dataset. In this CQA dataset, each question is associated with at most one correct answer and four negative answers [26].

SemEval-2016CQA: This is also a CQA dataset which is created from the Qatar Living Forums.² Each candidate answer is tagged with “Good”, “Bad” or “Potentially Useful”. We consider “Good” as positive and other tags as negative [13, 25].

SemEval-2017CQA: The training and validation dsata in this CQA dataset is same as SemEval-2016CQA. However, the test sets are different [20].

4.2 Training Parameters and Evaluation Metrics

We used both the cased and uncased models³ of BERT_{Large} and fine-tuned them for the pairwise sentence classification task [5]. The parameters of the BERT_{Large} model were: number of layers $\mathbf{L} = 24$, hidden size $\mathbf{H} = 1024$, number of self-attention heads $\mathbf{A} = 16$, feed-forward layer size $\mathbf{d}_{ff} = 4096$. For implementation, we used the Transformer library of Huggingface⁴ [30]. For training, we used cross entropy loss function to calculate the loss and utilized Adam as the optimizer. We set the batch size to 16 and ran 2 epochs with learning rate being set to 2×10^{-5} . We selected

¹ <https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=10>.

² <https://www.qatarliving.com/forum>.

³ https://huggingface.co/transformers/pretrained_models.html.

⁴ <https://github.com/huggingface/transformers>.

Table 3 Performance comparisons with the recent progress. Here, ‘FT’ denotes ‘Fine Tuning’

Model	QA datasets				CQA datasets					
	TREC-QA		WikiQA		YahooCQA		SemEval’16		SemEval’17	
	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
Kamath et al. [9]	0.852	0.891	–	–	–	–	–	–	–	–
Sha et al. [25]	–	–	0.746	0.758	–	–	0.801	0.872	–	–
Tay et al. [27]	–	–	–	–	–	0.801	–	–	–	–
Nakov et al. [20]	–	–	–	–	–	–	–	–	0.884	0.928
<i>BERT</i> _{Large(Cased)} FT	0.934	0.966	0.842	0.856	0.946	0.946	0.841	0.894	0.908	0.934
<i>BERT</i> _{Large(Uncased)} FT	0.917	0.947	0.843	0.857	0.951	0.951	0.866	0.927	0.921	0.963

the model for evaluation which performed the best in the validation set. To evaluate our models, we used the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR) as the evaluation metrics.

5 Results and Analyses

To evaluate the performance of fine-tuning the BERT model in the answer selection datasets, we compare its performance with various state-of-the-art models [9, 20, 25, 27]. We also conduct ablation studies to further investigate the effectiveness of fine-tuning. To note that, we pre-processed all datasets into the lower-cased format and evaluated with both the cased and uncased versions of the BERT model.

5.1 Performance Comparisons

We show the results of our models in Table 3. We find that in comparison to the prior work in the TREC-QA dataset, the fine-tuned *BERT*_{Large(Cased)} model performs the best and outperforms the previous state-of-the-art [9] with an improvement of 9.6% in terms of MAP and an improvement of 8.4% in terms of MRR. However, in the WikiQA dataset, the uncased version performs the best in terms of both MAP and MRR. More specifically, *BERT*_{Large(Uncased)} model improves the performance by 13% in terms of MAP and 13.1% in terms of MRR compared to the previous state-of-the-art [25] in the WikiQA dataset.

In the CQA datasets, we again observe that both models outperform the prior work. In terms of MRR, we find that the *BERT*_{Large(Uncased)} model outperforms [27],

Table 4 Performance comparisons based on the Ablation Test. Here, ‘FT’ denotes ‘Fine Tuning’

Model	QA datasets				CQA datasets					
	TREC-QA		WikiQA		YahooCQA		SemEval’16		SemEval’17	
	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
<i>BERT_{Large(Uncased)}</i> FT	0.917	0.947	0.843	0.857	0.951	0.951	0.866	0.927	0.921	0.963
<i>Without FT</i>	0.405	0.476	0.566	0.571	0.436	0.436	0.604	0.670	0.698	0.757

[25], and [20] with an improvement of 18.7%, 6.3%, and 3.8% in the YahooCQA, SemEval-2016CQA, and SemEval-2017CQA datasets respectively.

While comparing between the cased model and the uncased model, we find that even though the cased model outperforms the uncased model in the TREC-QA dataset, it fails to outperform the uncased model in other datasets. To be noted that, the cased model still provides competitive performance in comparison to the uncased model in all five datasets. In order to better analyze the performance of these two models, we conduct significant tests. Based on the paired t-test, we find that the performance difference between the two models is **not statistically significant** ($p \leq 0.05$). This indicates that the cased version of the fine-tuned BERT model is robust in scenarios when the datasets do not contain any casing information.

5.2 Ablation Studies

We perform ablation test to investigate the effectiveness of our approach of fine-tuning the BERT model. For the ablation test, we excluded fine-tuning and only used the feature-based embeddings generated from the pre-trained BERT_{Large (Uncased)} model. In our ablation study, we used all five datasets to compare the performance. From the ablation test (see Table 4), we find that removing fine tuning from BERT decreases the performance by 55.8, 32.9, 54.2, 30.3, and 24.2% in terms of MAP in the TREC-QA, WikiQA, YahooCQA, SemEval-2016CQA, and SemEval-2017CQA datasets respectively. The deterioration here without fine-tuning is **statistically significant** based on paired t-test ($p \leq 0.05$).

6 Conclusions and Future Work

In this paper, we adopt the pre-trained BERT model and fine-tune it for the answer selection task in five answer selection datasets. We observe that fine-tuning the BERT model for answer selection is very effective and find that it outperforms all the RNN-based models used previously for such tasks. In addition, we evaluate the

effectiveness of the cased version of the BERT model in scenarios when the casing information is not present in the target dataset and demonstrate that the cased model provides almost similar performance compare to the uncased model. We further investigate the effectiveness of fine-tuning the BERT model by conducting ablation studies and observe that fine-tuning significantly improves the performance for the answer selection task.

In the future, we will investigate the performance of different transformer-based models [13] on other tasks, such as information retrieval applications [6–8, 33], sentiment analysis [15, 16, 34], learning from imbalanced datasets [14], query-focused abstractive text summarization [12], and automatic chart question answering [10].

Acknowledgements This research is supported by the Natural Sciences & Engineering Research Council (NSERC) of Canada and an ORF-RE (Ontario Research Fund-Research Excellence) award in BRAIN Alliance. We thank anonymous reviewers for their thorough review comments on this paper and acknowledge Compute Canada for providing us with the computing resources. We also thank Dr. Qin Chen for helping us with the experiments.

References

1. Bari, M.S., Joty, S., Jwalapuram, P.: Zero-Resource Cross-Lingual Named Entity Recognition. arXiv preprint [arXiv:1911.09812](https://arxiv.org/abs/1911.09812) (2019)
2. Chen, Q., Hu, Q., Huang, J.X., He, L.: CA-RNN: Using context-aligned recurrent neural networks for modeling sentence similarity. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
3. Chen, Q., Hu, Q., Huang, J.X., He, L.: CAN: Enhancing sentence similarity modeling with collaborative and adversarial network. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 815–824 (2018)
4. Chen, Q., Hu, Q., Huang, J.X., He, L., An, W.: Enhancing recurrent neural networks with positional attention for question answering. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 993–996 (2017)
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)
6. Huang, X., Hu, Q.: A Bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 307–314 (2009)
7. Huang, X., Peng, F., Schuurmans, D., Cercone, N., Robertson, S.E.: Applying machine learning to text segmentation for information retrieval. *Inf. Retrieval* **6**(3–4), 333–362 (2003)
8. Huang, X., Zhong, M., Si, L.: York University at TREC 2005: genomics track. In: Proceedings of the Fourteenth Text Retrieval Conference, TREC (2005)
9. Kamath, S., Grau, B., Ma, Y.: Predicting and integrating expected answer types into a simple recurrent neural network model for answer sentence selection. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (2019)
10. Kim, D.H., Hoque, E., Agrawala, M.: Answering questions about charts and generating visual explanations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020)

11. Laskar, M.T.R., Hoque, E., Huang, J.: Utilizing bidirectional encoder representations from transformers for answer selection task. In: Proceedings of the V AMMCS International Conference: Extended Abstract, p. 221, (2019)
12. Laskar, M.T.R., Hoque, E., Huang, J.: Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In: Canadian Conference on Artificial Intelligence, pp. 342–348 (2020)
13. Laskar, M.T.R., Huang, X., Hoque, E.: Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 5505–5514 (2020)
14. Liu, Y., An, A., Huang, X.: Boosting prediction accuracy on imbalanced datasets with SVM ensembles. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, pp. 107–118 (2006)
15. Liu, Y., Huang, X., An, A., Yu, X.: ARSA: a sentiment-aware model for predicting sales performance using blogs. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 607–614 (2007)
16. Liu, Y., Huang, X., An, A., Yu, X.: Modeling and predicting the helpfulness of online reviews. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 443–452 (2008)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
18. Mayhew, S., Gupta, N., Roth, D.: Robust named entity recognition with truecasing pretraining. arXiv preprint [arXiv:1912.07095](https://arxiv.org/abs/1912.07095) (2019)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
20. Nakov, P., Hoogeveen, D., Márquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: Semeval-2017 task 3: Community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 27–48 (2017)
21. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
22. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237 (2018)
23. Rao, J., Liu, L., Tay, Y., Yang, W., Shi, P., Lin, J.: Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5373–5384 (2019)
24. Severyn, A., Moschitti, A.: Automatic feature engineering for answer selection and extraction. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 458–467 (2013)
25. Sha, L., Zhang, X., Qian, F., Chang, B., Sui, Z.: A multi-view fusion neural network for answer selection. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
26. Tay, Y., Phan, M.C., Tuan, L.A., Hui, S.C.: Learning to rank question answer pairs with holographic dual LSTM architecture. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 695–704 (2017)
27. Tay, Y., Tuan, L.A., Hui, S.C.: Hyperbolic representation learning for fast and efficient neural question answering. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 583–591 (2018)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

29. Wang, M., Smith, N.A., Mitamura, T.: What is the jeopardy model? a quasi-synchronous grammar for qa. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007)
30. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al.: Transformers: State-of-the-art natural language processing. arXiv preprint [arXiv:1910.03771](https://arxiv.org/abs/1910.03771) (2019)
31. Yang, Y., Yih, W.-t., Meek, C.: WikiQA: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018 (2015)
32. Yih, W.-t., Chang, M.-W., Meek, C., Pastusiak, A.: Question answering using enhanced lexical semantic models. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1744–1753 (2013)
33. Yin, X., Huang, J.X., Li, Z., Zhou, X.: A survival modeling approach to biomedical search result diversification using Wikipedia. *IEEE Trans. Knowl. Data Eng.* **25**(6), 1201–1212 (2013)
34. Yu, X., Liu, Y., Huang, X., An, A.: Mining online reviews for predicting sales performance: a case study in the movie domain. *IEEE Trans. Knowl. Data Eng.* **24**(4), 720–734 (2012)
35. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)

Calibration and Analysis of Structural Credit Risk Models with Occupation Time



Malhar M. Mukhopadhyay and Roman N. Makarov

Abstract Credit risk is concerned with analyzing financial losses occurring due to changes in the credit quality of a firm. A rare occurrence of such sort is the *default event* that often leads to bankruptcy or liquidation, resulting in large financial losses to investors. In structural credit risk models, the asset value is compared to firm's liabilities at any time, and the default event occurs when the asset value falls dangerously low. In this paper, we consider a structural credit risk model based on *occupation time*, which is defined as the time the firm asset value V spends below a default barrier. Here, we assume the geometric Brownian motion dynamics for V . Liquidation is declared if the firm value drops below a liquidation barrier or if it spends too much time below the default barrier. The main purpose of this paper is to calibrate the occupation-time model parameters using default probabilities derived from *Credit Default Swaps* (CDS) spreads available through Bloomberg Finance L.P.. This is done by applying the non-linear least-squares method. We also compare the occupation-time model with another well-known structural model of credit risk, namely, the Black–Cox model (1976).

Keywords Credit risk · Structural model · Liquidation · Default probability · Occupation time · Black—Cox model · Credit default swap

1 Introduction

Credit events are constantly present in the lives of people, financial institutions, and even countries. In short, credit risk is the uncertainty arising from potential default of an economic agent to another economic agent. A more formal definition of credit risk is as follows [1, 5]. *Credit Risk (Default Risk) is the potential loss arising from the*

M. M. Mukhopadhyay (✉) · R. N. Makarov
Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: mukh3990@mylaurier.ca

R. N. Makarov
e-mail: rmakarov@wlu.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_64

705

default of an economic agent to meet its contractual obligations in a pre-established period of time.

Under structural models of credit risk, a default event is deemed to occur for a firm when its assets reach a sufficiently low level compared to its liabilities. These models require strong assumptions on the dynamics of the firm's asset, its debt and how its capital is structured. In particular, we have

$$\text{Assets} = \text{Equity} + \text{Liabilities}$$

where Assets (V_t), Liabilities (B_t) and Equity (E_t) are considered time-dependent processes. Merton [9] modelled firm's assets through dynamics applied earlier by Black and Scholes. In the Merton model, the default may only be declared at maturity. The Black–Cox model [2] states that a default occurs if the company's assets have fallen below some predetermined default barrier after a certain period of time has passed. Both these models assume that there is no temporal separation between liquidation and default. The occupation time model [7] aims to provide a solution to this issue.

Assume that the firm's value follows a geometric Brownian motion under the risk-neutral measure \mathbb{Q} as given by the following SDE:

$$dV_t = rV_t dt + \sigma V_t dW_t. \quad (1)$$

In the Black–Cox model, default occurs at the first time the firm's value drops below a certain time-dependent barrier,

$$B(t) := Be^{-\gamma(T-t)} = B_0e^{\gamma t}, \quad 0 < t \leq T, \quad (2)$$

where γ is the exponential coefficient, which is comparable to the yield of a zero-coupon bond with face value B , and $B_0 = e^{-\gamma T}$. The default time in the Black–Cox model is then given by a solution to the following first-passage time problem:

$$\tau_B(V) = \inf\{t > 0 : V_t \leq B(t)\},$$

where $B_0 < V_0$.

2 Occupation-Time Structural Model

The occupation time is defined as the cumulative time spent by a process under a specific barrier. In our model, the cumulative time spent by the asset value process V under the default barrier B is

$$\mathcal{A}_t \equiv \mathcal{A}_t(V, B) = \int_0^t \mathbb{I}_{\{V(u) \leq B(u)\}} du. \quad (3)$$

We also introduce a predetermined threshold parameter $\nu > 0$. If the occupation time \mathcal{A}_t exceeds the threshold parameter, then the firm defaults at time

$$\tau_\nu = \inf\{t \geq 0 : \mathcal{A}_t \geq \nu\}. \tag{4}$$

We can also think of ν as the grace period granted by the bankruptcy court.

Let us also introduce another barrier given by $L(t) = L_0 e^{\gamma t}$, where $0 < L_0 < B_0 < V_0$. We call this barrier the *liquidation barrier*. If the firm's value drops below the liquidation barrier, the firm defaults even if the occupation time has exceeded the threshold parameter. That is, it is the first-hitting time for the process V with respect to the barrier L :

$$\tau_L \equiv \tau_L(V) = \inf\{t \geq 0 : V_t \leq L(t)\}. \tag{5}$$

In this model there are now two default times: τ_ν , which is due to the occupation time, and τ_L , which is due to the liquidation barrier. The liquidation time T_{Liq} is given by

$$T_{Liq} = \min\{\tau_\nu, \tau_L\}. \tag{6}$$

For any $t \geq 0$, we have $\{\tau_\nu \geq t\} = \{\mathcal{A}_t \leq \nu\}$. Therefore, the risk-neutral probability of liquidation by time T is

$$\begin{aligned} \mathbb{Q}(T_{Liq} \leq T) &= \mathbb{Q}(\min\{\tau_\nu, \tau_L\} \leq T) = 1 - \mathbb{Q}(\min\{\tau_\nu, \tau_L\} > T) \\ &= 1 - \mathbb{Q}(\tau_\nu > T, \tau_L > T) = 1 - \mathbb{Q}(\mathcal{A}_T < \nu, \tau_L > T). \end{aligned} \tag{7}$$

Applying a change of variables and using the fact that both default and liquidation barriers are exponential functions of time with the same exponent, we can write

$$\{V_t \leq B(t)\} = \{Z_t \leq b\} \text{ and } \{V_t \leq L(t)\} = \{Z_t \leq l\},$$

where $Z_t = x + \mu t + W_t$ is Brownian motion with drift, and the parameters are:

$$x = \frac{\ln V_0}{\sigma}, \quad \mu = \frac{r - \gamma - \frac{\sigma^2}{2}}{\sigma}, \quad b = \frac{\ln B_0}{\sigma}, \quad l = \frac{\ln L_0}{\sigma}.$$

Clearly, $x > b > l$ holds.

The occupation time $\mathcal{A}_T(V, B)$ and the first-passage time $\tau_L(V)$ have the same law as $\mathcal{A}_T(Z, b)$ and $\tau_l(Z)$, respectively:

$$\{\mathcal{A}_T(V, B) \leq t\} = \{\mathcal{A}_T(Z, b) \leq t\} \text{ and } \{\tau_L(V) > T\} = \{\tau_l(Z) > T\}.$$

Thus, the joint probability distribution from (7) can now be written as

$$\mathbb{Q}(\mathcal{A}_T(V, B) < \nu; \tau_L > T) = \mathbb{Q}(\mathcal{A}_T(Z, b) < \nu; \tau_l > T). \tag{8}$$

The probability density function of the occupation time \mathcal{A} with the condition $\min_{0 \leq t \leq T} \{Z_t\} > l$ can be obtained by the Laplace transform method. We use the following result from Makarov (2016) [8].

Theorem 1 *The probability of liquidation before maturity T is given by*

$$\begin{aligned} \mathbb{Q}(T_{Liq} \leq T) &= e^{-\mu^2 T/2} \left[\mathcal{L}_\gamma^{-1}[G_1(\gamma); v] + \int_0^v \mathcal{L}_\gamma^{-1}[g_2(\gamma); s] ds \right] \\ &= \frac{e^{-\mu^2 T/2}}{2\pi i} \left[\int_{C-i\infty}^{C+i\infty} G_1(\gamma; T-v) e^{\gamma v} d\gamma + \int_0^v \int_{C-i\infty}^{C+i\infty} g_2(\gamma; T-t) e^{\gamma t} d\gamma dt \right], \end{aligned} \tag{9}$$

where the functions G_1 and g_2 are given by

$$\begin{aligned} G_1(\gamma; \tau) &= \frac{\mu - c\sqrt{2\gamma}}{(\gamma - \mu^2/2)\sqrt{2}} e^{-\mu(x-b)} \sum_{j=1}^3 A_j c_j e^{ac_j + c_j^2 \tau} \operatorname{Erfc} \left(\frac{a}{2\sqrt{\tau}} + c_j \sqrt{\tau} \right), \\ g_2(\gamma; \tau) &= \frac{\sqrt{\gamma} e^{-\mu(x-l)}}{(\gamma - \mu^2/2) \sinh(\sqrt{2\gamma}(b-l))} \left(\frac{1}{\sqrt{\pi\tau}} e^{-\frac{(x-b)^2}{2}\tau} - \sqrt{\gamma} c e^{\sqrt{2\gamma}(x-b)c + \gamma^2 c^2 \tau} \right. \\ &\quad \left. \times \operatorname{Erfc} \left(\frac{x-b}{\sqrt{2\tau}} + \sqrt{\gamma} c \tau \right) \right). \end{aligned}$$

and $a = \sqrt{2}(x-b)$, $c = \coth(\sqrt{2\gamma}(b-l))$, $c_1 = -\frac{\mu}{\sqrt{2}}$, $c_2 = \frac{\mu}{\sqrt{2}}$, $c_3 = \sqrt{\gamma}c$. Also,

$$A_1 = \frac{1}{\mu(\mu + \sqrt{2\gamma}c)}, \quad A_2 = \frac{1}{\mu(\mu - \sqrt{2\gamma}c)}, \quad A_3 = \frac{1}{\gamma c^2 - \mu^2/2}.$$

Here, Erfc denotes the complementary error function.

The derivation of Eq. (9) follows the general derivation as also provided by Hugonnier [4] for a standard Brownian motion (see also [10]).

3 Numerical Results

3.1 Calibration Process

In the Black–Cox model, we calibrate the parameters γ , B_0 and σ , whereas in the occupation-time model with a liquidation barrier, we calibrate the parameters γ , B_0 , L_0 , σ and v . Let $P_{0,t}^{Mkt}$ denote the market implied probability of default from 0 to t as obtained from CDS prices available through Bloomberg Finance L.P., and

$P_{0,t}^{Mdl}$ denote the estimated probability of default obtained through the model. The calibration process can be represented by the following diagrams:

$$\left. \begin{matrix} P_{0,0.5}^{Mkt} \\ P_{0,1}^{Mkt} \\ \vdots \\ P_{0,10}^{Mkt} \end{matrix} \right\} \rightarrow \begin{cases} dV_t = rV_t dt + \sigma V_t dW_t \\ B(t) = B_0 e^{\gamma t} \\ \text{model parameters: } \sigma_V, B_0, \gamma, \end{cases} \quad (10)$$

for the Black–Cox model, and

$$\left. \begin{matrix} P_{0,0.5}^{Mkt} \\ P_{0,1}^{Mkt} \\ \vdots \\ P_{0,10}^{Mkt} \end{matrix} \right\} \rightarrow \begin{cases} dV_t = rV_t dt + \sigma V_t dW_t \\ B(t) = B_0 e^{\gamma t} \\ L(t) = L_0 e^{\gamma t} \\ \text{model parameters: } \sigma_V, B_0, \gamma, L_0, \nu, \end{cases} \quad (11)$$

for the occupation-time model.

We attempt to solve the following minimization problem:

$$\Theta := \operatorname{argmin} \sum_{t=0.5}^T w_t \left(P_{0,t}^{Mdl} - P_{0,t}^{Mkt} \right)^2 \quad (12)$$

where $\Theta := \{\sigma_V, B_0, \gamma, L_0, \nu\}$ for the occupation-time model with a liquidation barrier.

The weights $\{w_t\}$ are selected to achieve good curve fitting results. We keep the weights consistent with regards to the following.

- Companies within the same Moody’s rating tier have the same weights. If during the calibration process weights are changed for a certain firm, they are changed for all firms within the same Moody’s tier.
- We calculate parameters for the Black–Cox and the occupation-time models using the same weights for consistency.

The market implied probability of default can be estimated using CDS spreads that are obtained through Bloomberg Finance L.P.. We use the following relation [3, 5, 6]:

$$P_{0,T}^{Mkt} = \mathbb{Q}(\tau \leq T) = 1 - \exp\left(\frac{-s(T)T}{LGD}\right), \quad (13)$$

where T is the tenor, $s(T)$ is the corresponding CDS spread, and LGD is the loss given default. For simplicity we set the $LGD = 0.6$.

Using the expressions in (12) and (13), the calibration process is then carried out as follows.

- The initial values of $V_0 = 100$, $B_0 = 90$, $L_0 = 80$, $\sigma = 0.2$, $r = 0.05$, $\gamma = r$ are chosen for calibrating the Black–Cox model parameters.
- Using the parameters obtained for the Black–Cox models for $\gamma = \gamma_{BC}$, $\sigma = \sigma_{BC}$, $B_0 = B_0^{BC}$, we calibrate the model parameters for the occupation-time model.
- If the occupation-time model fits the curve well, the process is ended and the weights are then used for another firm within the same Moody’s rating tier. Otherwise, the calibration process is run again with changed weights depending on how well the curve was fitted for different tenors.
- The calibration process was done sequentially for tenors $0.5 \leq t \leq T$ where the maturity $T \rightarrow 10$. The parameters obtained from one calibration cycle were used as starting values for the next cycle with a larger maturity.

3.2 Case Studies

3.2.1 Corporate CDS Prices

For the purpose of this paper we present the results for companies in different Moody’s Ratings Tier, namely, Johnson & Johnson [JNJ], J.P. Morgan Chase & Co. [JPMCC], Kraft–Heinz Co. [KHC] and Ford Motor Co. [F] (Fig. 1, Tables 1 and 2).

3.2.2 Turkish Sovereign CDS

On August 15, 2018, the Turkish Sovereign CDS curve inverted, indicating that the short term insurance against default is more than the long term protection. CDS inversions occur when investors are worried about collapse in the short term and are inclined more towards purchasing long term security. In the case of Turkish Sovereign CDS inversion, it was deemed that investors were worried about the collapse of the relevant Turkish USD Sovereign bonds with short term maturities. Figure 2a compares the calibrated implied probability with the market data. As seen from Table 3, the value of γ is negative. Figure 2b shows the inverted curve.

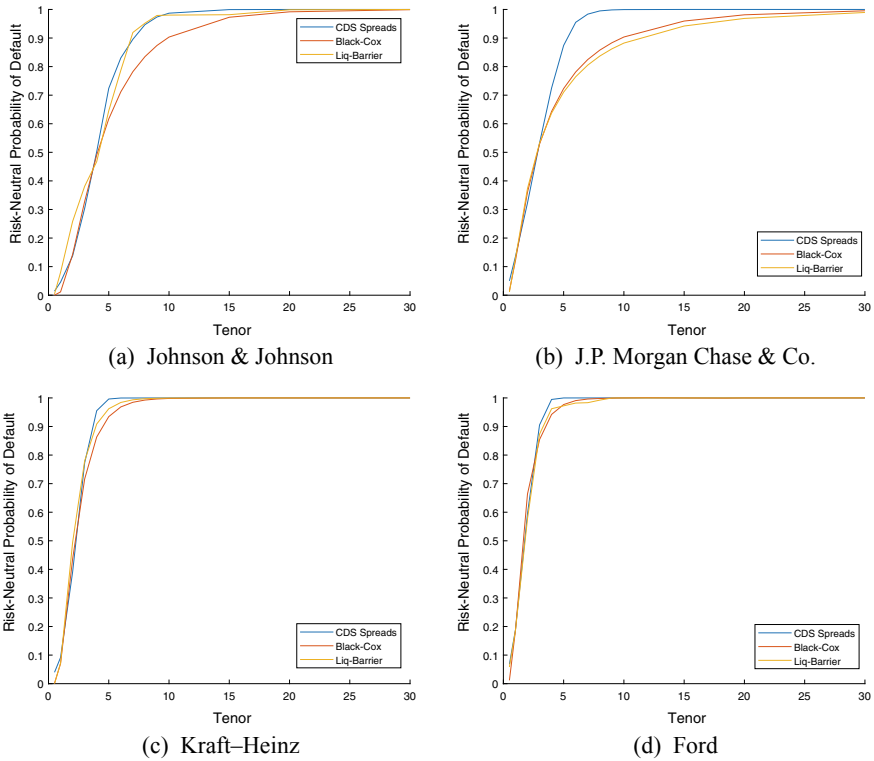


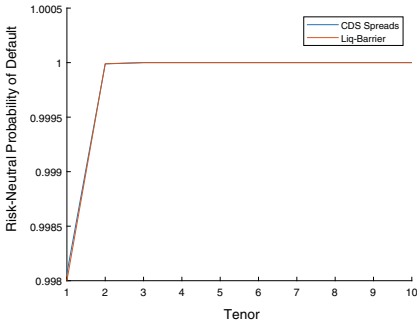
Fig. 1 Implied risk-neutral probabilities of default for four companies

Table 1 Parameters for the Black-Cox Model: four companies

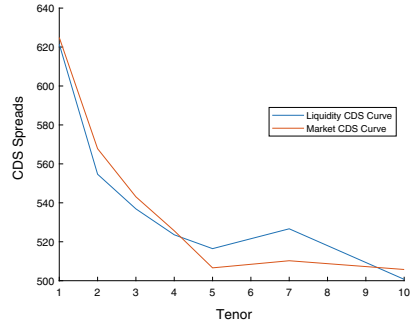
CDS	B_0	σ	γ
JNJ	2.1110	0.9956	-0.0044
JPMCC	64.1751	0.2329	0.123
KHC	74.8282	0.1047	0.1587
F	76.3494	0.925	0.1672

Table 2 Parameters for the occupation-time model: four companies

CDS	ν	B_0	L_0	σ	γ
JNJ	0.2324	55.3446	36.3215	0.4502	0.0012
JPMCC	0.1593	72.9971	49.2994	0.2281	0.1052
KHC	0.1362	73.072	39.5663	0.1307	0.1993
F	0.1445	72.3683	42.2334	0.1239	0.1728



(a) Implied risk-neutral probability of default.



(b) Real-world data vs. model curve.

Fig. 2 Turkey Sovereign CDS Curve: calibration of the occupation-time model with a liquidation barrier

Table 3 Parameters for the occupation-time model: Turkish Sovereign CDS

ν	B_0	L_0	σ	γ
0.0002	9.4601	0.2063	6.2601	-2.5724

4 Conclusion

The main objective of this paper was to calibrate the parameters for the occupation time model with default and liquidation barriers. We looked at firms in different tiers of Moody’s ratings and observed that across various tiers the calibrated parameters had similar properties. None of the parameters single-handedly got affected by CDS prices, and, in fact, all parameters changed in the implied direction as we went up or down the Moody’s ratings tiers. The occupation-time model replicated the CDS implied default probability better than the Black–Cox model, which confirmed the flexibility due to more parameters available for calibration. In the case of the Turkish CDS’s, the implied CDS spread obtained from model parameters matched perfectly with the CDS curve obtained from the market, even though there was a CDS spread inversion. The occupation-time model also allows for the temporal separation of default and liquidation and can be used in pricing market derivatives that consider liquidation.

Acknowledgements R. Makarov wishes to acknowledge the support of the NSERC Discovery Grant program.

References

1. Bieleck, T.R., Rutkowski, M.: *Credit Risk: Modeling, Valuation and Hedging*. Springer Science & Business Media (2013)
2. Black, F., Cox, J.C.: Valuing corporate securities: some effects of bond indenture provisions. *J. Finance* **31**(2), 351–367 (1976)
3. Brigo, D., Tarengi, M.: Credit default swap calibration and equity swap valuation under counterparty risk with a tractable structural model. Available at SSRN 581302 (2004)
4. Hugonnier, Julien-N: The Feynman-Kac formula and pricing occupation time derivatives. *Int. J. Theor. Appl. Financ.* **2**(02), 153–178 (1999)
5. Hull, J.: *Options, futures and other derivatives*. Prentice Hall, Upper Saddle River, NJ (2009)
6. Hull, J.C., Predescu, M., White, A.: Bond prices, default probabilities and risk premiums. In: *Default Probabilities and Risk Premiums* (2005)
7. Makarov, R., Metzler, A., Ni, Z.: Modelling default risk with occupation times. *Finan. Res. Lett.* **13**, 54–65 (2015)
8. Makarov, R.N.: Modeling liquidation risk with occupation times. *Int. J. Financ. Eng.* **3**(4), 1650028 (2016)
9. Merton, R.C.: On the pricing of corporate debt: the risk structure of interest rates. *J. Financ.* **29**(2), 449–470 (1974)
10. Zhu, Song-Ping; Chen, Wen-Ting: Pricing Parisian and Parasian options analytically. *J. Econ. Dyn. Control* **37**(4), 875–896 (2013)

Prediction Intervals of Machine Learning Models for Taxi Trip Length



Ella Morgan, Ryan Zhou, and Wenying Feng

Abstract Errors are always present in predictions produced by machine learning models. Producing a quantitative estimate of the uncertainty in a model's output is crucial for many fields, especially those where predictive models drive important decisions. In this paper we discuss two methods for producing prediction intervals for neural network, random forest, and gradient boosted tree models. We then evaluate the prediction intervals produced by each algorithm by predicting the expected ride length for a NYC taxi trip dataset. We show that inductive conformal prediction produces the most reliable intervals for all machine learning models investigated.

Keywords Machine learning · Neural network · Gradient boosted tree · Random forest · Prediction interval

1 Introduction

No model is correct all the time. This is a reality for all machine learning algorithms, from simple linear regressions to deep neural networks. Recognizing when a model may produce inaccurate predictions is of increasing importance, especially with the growing usage of predictive models in the real world. As such it is helpful to obtain, along with the point estimate produced by the model, a quantitative measurement of a model's uncertainty in the form of a variance or a prediction interval.

Methods for producing prediction intervals for a learning algorithm such as linear regression have been well studied. However, these methods are not easily generalizable to more complex models such as those used in machine learning. We

E. Morgan · R. Zhou · W. Feng (✉)
Trent University, Peterborough, ON, Canada
e-mail: wfeng@trentu.ca

E. Morgan
e-mail: ellamorgan@trentu.ca

R. Zhou
e-mail: ryanzhou@trentu.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_65

715

will present two distinct approaches—decomposition into model variance and inherent noise, and inductive conformal prediction—for producing prediction intervals in common machine learning applications. To our knowledge, these methods have not been directly compared in the literature. Our primary contribution with this paper is an empirical comparison of these methods applied to multiple machine learning algorithms. To this end, we develop a method of estimating model uncertainty on tree-based algorithms as well as an improvement on the inherent noise calculation by dividing the calibration set into subsets based on target values.

We evaluate these methods on a dataset consisting of taxi trips in New York City, using machine learning algorithms to predict the duration of a trip. This problem is ideal for showcasing the techniques for a few reasons. Taxi trip duration can be extremely difficult to predict as it relies on unpredictable external factors which are difficult to quantify. Traffic, weather, and road conditions can change rapidly and unpredictably, varying the actual duration of a trip by a large amount and ensuring that errors will be inevitable to some degree. In addition, measurable factors such as the starting and ending locations and the time of day do not have simple relationships with the trip duration, making this a complex problem ideal for learning efficiency evaluation.

The algorithms we use to generate predictions are random forest, gradient boosting, and neural networks. Random forest [4] is an ensemble method that aggregates the predictions of multiple decision trees to produce a single final prediction. Each tree is trained on a subset of the training data produced by drawing with replacement and choosing a random subset of features for each split in the tree. The prediction is then the average prediction between all trees.

Gradient boosting [6] is a tree-based ensemble method closely related to random forest. However, rather than training all trees independently, it creates the trees iteratively and attempts to improve on the ensemble of existing trees by using gradient descent over the loss function. Each individual tree can be used as a weak estimator as well, a trait which we use for uncertainty estimation.

Neural networks consist of many nodes or neurons, each of which produces an output that is a linear combination of its inputs modified by a nonlinear activation function. Neural network models link together many neurons in multiple layers in order to learn complex relationships between input and output. Regularization is often used on neural networks to alleviate overfitting; one such regularization technique is dropout, which randomly mutes the outputs from a certain fraction of neurons. This approximates the training of a large number of similar neural networks with shared weights [12], forcing the network to generalize better by relying less on specific weights. As we describe in the next section, this approximation to ensemble behaviour can also be used for uncertainty estimation.

2 Prediction Interval Algorithms

We provide two techniques for constructing a prediction interval. The first involves decomposing the total uncertainty into two components, one captures uncertainty within the model and the other describes the uncertainty in the input data. The second technique, conformal prediction, approaches the problem from a different angle with attempting to evaluate the difficulty of each input data point by comparing it to other known points.

2.1 Decomposition into Model Variance and Inherent Noise

Total prediction uncertainty is a combination of model uncertainty, denoted by η_1^2 , and inherent noise in the data, denoted by η_2^2 [13]. Model uncertainty is specific to the underlying trained model, and results from uncertainty in the model parameter estimates. We will first discuss inherent noise, and then discuss different methods for evaluating model uncertainty.

The inherent noise η_2^2 is calculated by finding the mean squared error on a separate validation set as follows:

$$\eta_2^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \tag{1}$$

where y_i represents the target value and \hat{y}_i is its corresponding prediction. This method produces one η_2^2 for the full validation set and this single value is used for all prediction intervals. We propose a modification to this method, where instead the validation set are sorted based on the predicted outputs \hat{y}_i and split into n separate quantiles. The inherent noise is then evaluated separately for each quantile, so that each η_2^2 corresponds to a different range of prediction outputs. At inference time, we use η_2^2 of the quantile which contains the test prediction.

The two variances, η_1^2 and η_2^2 , are then combined into a total uncertainty measure in the following manner:

$$\eta = \sqrt{\eta_1^2 + \eta_2^2}. \tag{2}$$

We now describe some methods for calculating η_1^2 , the model uncertainty. These methods are specific to the model used, whereas inherent noise is independent of the model and depends only on the data.

Variance Estimation Through Monte Carlo Dropout

As outlined earlier, dropout is a regularization technique used on neural networks to prevent overfitting. This is done by muting a certain fraction of neurons at random at each training epoch, effectively changing the structure of the neural network slightly each time [12]. While this prevents the network from being overly reliant on specific nodes by forcing it to make predictions when the nodes in question are dropped out,

it can also be seen as performing an averaging of the weights over the ensemble of possible subnetworks. As such, the final dropout-trained network can be viewed as an approximation of an ensemble prediction.

Following the method described by Gal in [7], dropout can also be used during inference time to effectively reconstruct the ensemble and gather predictions from random subnetworks, a technique known as Monte Carlo dropout. By making a large number of predictions with dropout, we are able to sample from the distribution of the ensemble predictions and estimate their variance as follows:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B \hat{y}_b, \quad (3)$$

$$\eta_1^2 = \frac{1}{B} \sum_{b=1}^B (\hat{y}_b - \hat{y})^2, \quad (4)$$

where \hat{y}_b represents a single prediction made with dropout and B is the duplication of the experiments, or the number of times the prediction procedure was repeated. The average of all predictions is returned as the point prediction \hat{y} , and the variance is found by Eq. (4). This variance is the model uncertainty, representing the disagreement in the predictions of subnetworks produced by dropout, despite having all been trained on the same dataset.

Variance Estimation for Tree Based Estimators

Although dropout regularization is a neural network technique and not used with tree based models, random forest and gradient boosted trees are innately ensemble methods. Therefore, we propose to calculate the model uncertainty using the same method as the one described above, by finding the variance of the predictions produced by submodels in the ensemble. With \hat{y}_b in Eq. (3) now being the predictions from individual trees, we similarly calculate the model uncertainty for the tree based prediction with Eq. (4). This method can also be used with other ensemble algorithms beyond those described in this paper.

Variance Estimation Through Proper Scoring

It is possible to take a different approach, by training a neural network to predict its own variance. We modify the neural network to output a probability distribution instead of a single value; as described in [9], this amounts to outputting the mean and variance for a Gaussian distribution. The loss function to be minimized then is the negative log-likelihood criterion:

$$-\log p_{\theta}(y_n|x_n) = \frac{\log \sigma_{\theta}^2(x_n)}{2} + \frac{(y - \mu_{\theta}(x_n))^2}{2\sigma_{\theta}^2(x_n)}, \quad (5)$$

where μ_{θ} and σ_{θ}^2 are the outputted mean and variance respectively. By modifying the neural network in this manner, the mean and variance can be estimated without producing multiple predictions as with Monte Carlo dropout, at the cost of increasing

the number of parameters in the neural network. The model uncertainty produced by this method η_1^2 is then the value σ_θ^2 predicted by the network: $\eta_1^2 = \sigma_\theta^2(x)$.

2.2 *K* Nearest Neighbors Inductive Conformal Prediction

Conformal prediction classifies data points based on a nonconformity measure—or intuitively, how “strange” each data point is for the underlying model—based on the error the model produces when labeling that point. This is measured using a reserved validation set, as shown in Algorithm 1. A calibration score is chosen by ordering the nonconformity measures and finding the value which captures all scores up to a given significance level. As each score is an explicit function of error, this can be directly translated into a prediction interval.

Newer formulations [10] of conformal prediction also employ normalization using a difficulty measure, which is used to adjust the sizes of the uncertainty intervals based on the estimated difficulty of the data point. For this study we use the average error, weighted by distance, of the K nearest neighbours as the difficulty estimate. This version of the metric is described in further detail in [3].

Algorithm 1 kNN ICP

Input: Data x^* , validation set x' , prediction algorithm $h(\cdot)$, kNN algorithm $g(\cdot)$, number of nearest neighbors k , significance level δ , sensitivity parameter β

Output: Prediction interval \hat{Y}_i^δ

// prediction

$\hat{y} \leftarrow h(x^*)$

$\hat{y}' \leftarrow h(x')$

for x'_i in validation set x' **do**

// find K nearest neighbors

$\{x'_1, \dots, x'_k\} = g(x'_i, x', k)$

for x'_j in $\{x'_1, \dots, x'_k\}$ **do**

// euclidean distance

$d_j = d(x'_i, x'_j)$

$o_j = |\hat{y}'_j - y'_j|$

end for

// difficulty measurement, ϵ is a small value that prevents division by zero

$\mu_i = \frac{\sum_{j=1}^k o_j / (d_j + \epsilon)}{\sum_{j=1}^k 1 / (d_j + \epsilon)}$

// nonconformity measure

$\alpha_i = \frac{o_i}{\mu_i + \beta}$

end for

sort $\{\alpha_1, \dots, \alpha_m\}$ incrementally

$\alpha_\delta = \alpha_{\lfloor \delta(m+1) \rfloor}$

$\hat{Y}_i^\delta = \hat{y}_i \pm \alpha_\delta(\mu_i + \beta)$

return \hat{Y}_i^δ

3 Computational Models for Empirical Evaluation

In this section we describe the setup for an empirical evaluation of the different algorithms. We report the results from the tests in the following section.

The dataset used for the experiments was obtained from the New York City (NYC) Taxi and Limousine Commission website <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>, for the month of August, 2016. In order to prevent data leakage, we used only features which would be known before a taxi trip, including pickup latitude and longitude, drop off latitude and longitude, distance travelled (which could be estimated with the routing software used by taxi drivers), and the day and hour of pickup.

The dataset was augmented with features such as the average taxi speed that hour, the pickup and drop off neighborhood (by grouping locations based on density), the number of taxi rides that hour, the direction of travel, and whether the taxi ride started or ended at an airport. Records were removed if they failed to meet any of the following criteria: a trip time between 10s and 20h, a trip distance greater than 0, an origin and destination within NYC and with valid coordinates, and an average trip speed (time/distance) under 100km/h.

The models in our study include (1) random forest; (2) gradient boosting; (3) a neural network with a dropout layer; and (4) a neural network using a proper scoring method. The variance and inherent noise methods were used to construct intervals for predictions for all four models. The K nearest neighbours inductive conformal prediction method (Algorithm 1) was used to construct intervals for predictions made by models (1), (2) and a standard neural network. For these models we set $\epsilon = 0.001$ and $\beta = 0.01$. In total 7 computational models were evaluated based on their prediction intervals.

Random forest and gradient boosting were implemented using the package scikit-learn [11] in Python. For both models 1000 trees were built, and all other parameters were the defaults in scikit-learn. For both random forest and gradient boosting a single model was trained and there were one set of predictions for each, and then confidence intervals were found using two different methods described in Sect. 2.

All neural networks were implemented in Python using Keras [5] and Tensorflow [1]. Three different models were built, all consisting of a dense layer of 1000 neurons using a tanh activation. The first neural network implemented variance estimation through Monte Carlo dropout and incorporated a dropout layer active during both training and testing time. The second network used the proper scoring method to output the mean and variance of a Gaussian distribution. This network was trained using log-likelihood loss implemented as a custom loss function in Keras. The mean of the distribution was taken to be the model's prediction, and the variance was used as model uncertainty. The final neural network was a standard neural network used as the underlying model for kNN inductive conformal prediction.

To compare the accuracy of predictions for the learning models, the mean absolute error (MAE), root mean square error (RMSE), and symmetric mean absolute percentage error (SMAPE) scoring metrics are applied. One benefit for the RMSE is

that larger errors get penalized more and have a larger impact on the results, which helps to evaluate the quantity of outliers in the predictions. The SMAPE defined by (6) was used due to its popularity in the literature and ability to handle small values well [2, 8]:

$$\text{SMAPE} = \frac{100}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}. \tag{6}$$

In addition to these commonly applied metrics for comparisons within the same dataset, we compare the prediction intervals between datasets using two main criteria. First, we investigate whether the intervals capture the desired confidence level by finding the fraction of target values y in the testing set which fall inside the calculated prediction interval $[\hat{y}_{lower}, \hat{y}_{upper}]$, where \hat{y}_{upper} and \hat{y}_{lower} are the upper and lower limits of the interval, respectively. The percentage of records found within the interval, referred to as capture percentage (CP), is calculated as follows and should be close to the target confidence level as desired:

$$\text{CP} = 100 \times \frac{1}{N} \sum_{n=1}^N c_n, \tag{7}$$

$$c_n = \begin{cases} 1 & \hat{y}_{upper} \geq y \geq \hat{y}_{lower}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Next the relative sizes of the prediction intervals are compared. Given that the accuracy of the intervals remains around the intended percentage, we prefer methods which produce smaller intervals on average. To this end, we find the mean and median interval size. Using both the mean and median allows us an overview of the comparative sizes of the intervals while remaining robust to outliers.

4 Experimental Results

Table 1 compares the performance of the models based on their point predictions. The random forest model achieved the highest accuracy by all listed metrics while the gradient boosting model proved the least accurate. Of the neural networks, the standard neural network performed slightly better than other neural network models, indicating that dropout and the proper scoring methods affect accuracy by a small degree. However, the overall difference in performance between models was small and all models obtained comparable accuracy on this dataset.

Table 2 evaluates the performance for different methods of computing predictive regions. The results show that K nearest neighbours inductive conformal prediction (models 5–7) consistently outperformed variance and inherent noise methods (models 1–4) in both accuracy of coverage by Eq. (7) and interval size. The intervals

Table 1 Point prediction results

Model	MAE	RMSE	SMAPE
Random forest	143.14	218.24	18.89
Gradient boosting	153.24	230.55	20.39
Standard neural network	146.12	226.29	19.58
Neural network with dropout	148.47	232.53	19.40
Neural network with proper scoring	150.60	231.10	19.82

Table 2 Prediction interval results

Model	90% Interval			95% Interval			99% Interval		
	CP (%)	Mean	Median	CP (%)	Mean	Median	CP (%)	Mean	Median
1	98.0	459.5	414.4	98.9	547.5	493.8	99.6	719.6	649.0
2	94.7	390.1	315.4	97.1	464.9	375.8	98.9	611.0	494.0
3	95.5	388.0	366.9	97.6	462.3	437.1	99.0	607.6	574.5
4	94.3	371.3	338.5	96.8	442.4	403.4	98.7	581.4	530.1
5	89.8	327.2	313.9	94.8	443.1	425.0	98.9	766.2	735.1
6	89.6	346.8	333.7	94.8	466.6	449.0	98.8	785.1	755.4
7	89.5	301.1	264.8	94.8	385.0	338.6	98.9	616.2	541.8

The capture percentage (CP) column shows the percentage of intervals that contain the true value obtained by (7). The mean and median columns show the average sizes of the prediction intervals.

Model 1: Random forest with variance and inherent noise method for tree based estimators.

Model 2: Gradient boosting with variance and inherent noise method for tree based estimators.

Model 3: Neural network with variance and inherent noise method with Monte Carlo dropout variance.

Model 4: Neural network with variance and inherent noise method using a proper scoring method.

Model 5: Random forest with K nearest neighbours inductive conformal prediction.

Model 6: Gradient boosting with K nearest neighbours inductive conformal prediction.

Model 7: Neural network with K nearest neighbours inductive conformal prediction

obtained from conformal prediction better captured the target confidence level and as a result, prediction intervals produced by conformal prediction were also smaller due to the variance and inherent noise method consistently producing larger intervals than necessary. The variance and inherent noise method produced more conservative estimates which may be desirable if avoiding underestimates is critical, but the significantly larger intervals do not as accurately convey the model’s confidence in its prediction. For instance, at the 90% prediction interval level, the capture percentage (CP) for models 5–7 are all within 1% of the target CP, where models 1–4 are between 4 and 8% more than the desired CP. Similar results are observed for the 95 and 99% prediction intervals. When the interval size is also considered, model 7 produced the best result with the smallest mean and median interval width among all the models.

With the variance and inherent noise method, the gradient boosting and neural network models produced comparable intervals. Between the proper scoring and dropout neural networks, proper scoring was consistently better and produced smaller

intervals while still meeting the target capture percentage. The neural network with proper scoring was also significantly faster at generating intervals and predictions compared to the dropout method, as sampling a large number of predictions for the dropout network added considerable overhead.

Overall, the inductive conformal prediction was consistently able to achieve the target confidence level, independent of the underlying algorithm. This illustrates a key advantage of conformal prediction in that the algorithm can be treated as a black box, unlike variance decomposition which requires model-specific methods of estimating the model uncertainty. In addition, the inductive version of conformal prediction used in this paper only requires a relatively straightforward calculation of difficulty at inference time, unlike some implementations of variance estimation which may require a computationally expensive ensemble prediction.

5 Conclusion

In this paper, methods to produce prediction intervals for machine learning algorithms were evaluated. We compared two approaches to finding such intervals: a decomposition into model variance and inherent noise method, and an inductive conformal prediction method utilizing K nearest neighbours. These techniques were evaluated on three base models trained to predict taxi trip length. In terms of prediction accuracy we achieved comparable results on all models, with random forest having marginally better results. For prediction intervals we found that K nearest neighbors inductive conformal prediction outperforms the variance with inherent noise methods for all cases. It also has the significant advantage of being independent from the underlying algorithm, so it is compatible with and can be implemented alongside any prediction model.

As future work, theoretical analysis would provide better explanations to the experimental results. Testing with bigger datasets from other application areas will also enhance the conclusions.

Acknowledgements The project was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016)
2. Scott Armstrong, J.: Long-Range Forecasting: From Crystal Ball to Computer, 2nd edn. Wiley, New York (1985)
3. Boström, H., Linusson, H., Löfström, T., Johansson, U.: Accelerating difficulty estimation for conformal regression forests. *Ann. Math. Artif. Intell.* **81**(1–2), 125–144 (2017)

4. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
5. François, C. et al.: Keras. <https://keras.io> (2015)
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189–1232 (2001)
7. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *International Conference on Machine Learning*, pp. 1050–1059 (2016)
8. Makridakis, Spyros: Accuracy measures: theoretical and practical concerns. *Int. J. Forecast.* **9**(4), 527–529 (1993)
9. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 1, pp. 55–60. IEEE (1994)
10. Regression conformal prediction with nearest neighbours: Papadopoulos, Harris., Vovk, Vladimir, Gammerman, Alexander. *J. Artif. Intell. Res.* **40**, 815–840 (2011)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
13. Zhu, L., Laptev, N.: Deep and confident prediction for time series at uber. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 103–110. IEEE (2017)

The Cobb-Douglas Production Function Revisited



Roman G. Smirnov and Kunpeng Wang

Abstract Charles Cobb and Paul Douglas in 1928 used data from the US manufacturing sector for 1899–1922 to introduce what is known today the Cobb-Douglas production function that has been widely used in economic theory. We employ the R programming language to fit the formulas for the parameters of the Cobb-Douglas production function generated by the authors recently via the bi-Hamiltonian approach to the same data set utilized by Cobb and Douglas. We conclude that the formulas for the output elasticities and total factor productivity fit to the data quite well.

Keywords Data fitting · Cobb-Douglas production function · Bi-Hamiltonian approach · Dynamical systems and statistical methods · R Programming

1 Introduction

The study and applications of the Cobb-Douglas production function in the field of economic science have a long history. Recall that in 1928 Charles Cobb and Paul Douglas published their seminal paper [1] in which the authors established a relationship between the volume of physical production in American manufacturing from 1899 to 1922 and the corresponding changes in the amount of labor and capital that had been employed during the time period to turn out the said physical production. More specifically, the authors computed and expressed in logarithmic terms the index numbers of the fixed capital, total number of production workers employed in American manufacturing, and physical production in manufacturing. It was established that the curve for production lied approximately one-quarter of

R. G. Smirnov (✉) · K. Wang
Department of Mathematics and Statistics, Dalhousie University, 6316 Coburg Road,
PO BOX 15000, Halifax, Nova Scotia B3H 4R2, Canada
e-mail: Roman.Smirnov@dal.ca

K. Wang
e-mail: kunpengwang@dal.ca

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_66

725

the distance between the curves representing the corresponding changes in labor and capital. Accordingly, Cobb and Douglas adopted the following function (previously also used by Wicksteed and Wicksell) given by

$$Y = f(L, K) = AL^k K^{1-k}, \quad (1)$$

where Y , L , and K represent production, labor, and capital respectively, while A is total factor productivity. The authors used the method of least squares to find that for the value of $k = 3/4$ the estimated values of Y fairly well approximated the actual values for the actual production in American manufacturing from 1899 to 1922.

It took 20 more years of careful research and scrupulous study of different data before the economic community accepted the formula (1), although the research continued past the 1947 Douglas' presidential address given to the American Economics Association in Chicago that marked the overall acceptance of the results of the original research conducted in 1928 by Cobb and Douglas (see [3] for a historical review and more details) and is still being done [2]. Notably, the Cobb-Douglas aggregate production function is still being used to fit to data coming from different fields of study driven by growth in production (see, for example, Prajneshu [4]).

The next milestone in the development of the theory behind the Cobb-Douglas production function (1) that we wish to highlight in this paper is the research conducted by Ruzyo Sato [5] (see also Sato and Ramachandran [6] for more references and details) in which the author derived the Cobb-Douglas production function under the assumption of exponential growth in production, labor and capital, using some standard techniques from the Lie group and dynamical systems theories. Sato's results were further developed and extended recently by the authors in [7] under the assumption of logistic rather than exponential growth in production and factors (labor and capital). Under the assumptions specified the author derived in a straightforward matter the general form of the Cobb-Douglas function. More specifically, the function derived by Sato is of the following form:

$$Y = f(L, K) = AL^\alpha K^\beta, \quad (2)$$

where Y , L and K are as before, while α and β denote the corresponding elasticities of substitution. However, in order to assure that the elasticities of substitution α and β admitted economically accepted values for $\alpha, \beta > 0, \alpha + \beta = 1$ as in (1), Sato had to assume that the function in question was holothetic under two types of technical change simultaneously that assured the same form for the production function (2) as in the original paper by Cobb and Douglas [1].

Recently the authors have improved the result by Sato by employing the bi-Hamiltonian approach [8]. More specifically, it was shown that the exponential growth in the factors of production and production under some mild assumption leads to the same form of the Cobb-Douglas production function (2) without Sato's assumption of simultaneous holotheticity [5].

The main goal of this paper is to establish a link between the analytic approach to the problem of the derivation of the Cobb-Douglas production function presented in [8] and the original data studied by Cobb and Douglas in [1] by employing R Programming.

2 Theoretical Framework

In this section we briefly review the three approaches to the problem of the derivation of the Cobb-Douglas function outlined in the introduction.

First, Cobb and Douglas in [1] presented a comprehensive study of the elasticity of labor and capital and how their variations affected corresponding volume of production in American manufacturing from 1899 to 1922. In particular, they plotted the corresponding time series of production output (Day index of physical production), labor and capital on a logarithmic scale (see Chart I in [1]). Since we will use this data in what follows, let us first tabulate the index numbers of the industrial output in American manufacturing Y , fixed capital K , and total number of manual workers L on a logarithmic scale in the following table.

The authors demonstrated with the aid of the method of least squares that the above data presented in Table 1 is subject to the following formula:

$$Y = f(L, K) = 1.01L^{3/4}K^{1/4}, \tag{3}$$

which is a special case of the formula (2).

Next, recall Sato employed in [5] an analytic approach to derive the Cobb-Douglas function (2). Summed up briefly, his approach was based on the assumption that the production and the corresponding input factors (labor and capital) grew exponentially. Under this assumption the problem of the derivation of the Cobb-Douglas function comes down to solving the following partial differential equation:

$$X\varphi = aK \frac{\partial\varphi}{\partial K} + bL \frac{\partial\varphi}{\partial L} + cf \frac{\partial\varphi}{\partial f} = 0, \tag{4}$$

where $\varphi(K, L, f) = 0$, $\partial\varphi/\partial f \neq 0$ is a solution to (4). Solving the corresponding system of ordinary differential equations

$$\frac{dK}{aK} = \frac{dL}{bL} = \frac{df}{cf}, \tag{5}$$

using the method of characteristics, yields the function (2), where $\alpha = \alpha(a, b, c)$, $\beta = \beta(a, b, c)$. Unfortunately, the elasticity elements in this case do not attain economically meaningful values as in (1), because of the condition $\alpha\beta < 0$. To mitigate this problem Sato in [5] introduced the notion of the simultaneous holothenticity, which implied that a production function in question was holothetic under more than

Table 1 The time series data used by Charles Cobb and Paul Douglas in [1]

Year	Output Y	Capital K	Labour L
1899	4.605170	4.605170	4.605170
1900	4.615121	4.672829	4.653960
1901	4.718499	4.736198	4.700480
1902	4.804021	4.804021	4.770685
1903	4.820282	4.875197	4.812184
1904	4.804021	4.927254	4.753590
1905	4.962845	5.003946	4.828314
1906	5.023881	5.093750	4.890349
1907	5.017280	5.170484	4.927254
1908	4.836282	5.220356	4.795791
1909	5.043425	5.288267	4.941642
1910	5.068904	5.337538	4.969813
1911	5.030438	5.375278	4.976734
1912	5.176150	5.420535	5.023881
1913	5.214936	5.463832	5.036953
1914	5.129899	5.497168	5.003946
1915	5.241747	5.583469	5.036953
1916	5.416100	5.697093	5.204007
1917	5.424950	5.814131	5.278115
1918	5.407172	5.902633	5.298317
1919	5.384495	5.958425	5.262690
1920	5.442418	6.008813	5.262690
1921	5.187386	6.033086	4.990433
1922	5.480639	6.066108	5.081404

one type of technical change simultaneously. Economically, this assumption leads to a model with the aggregate production function described by exponential, say, growth in two different sectors of economy (or, two countries) rather than one. From the mathematical perspective, this model yields a production function which is an invariant of an integrable distribution of vector fields Δ on \mathbb{R}_+^2 , each representing a technical change determined by the formula (4) if both of them are determined by exponential growth. Indeed, consider the following two vector fields, for which a function $\varphi(K, L, f)$ is an invariant:

$$X_1\varphi = K \frac{\partial\varphi}{\partial K} + L \frac{\partial\varphi}{\partial L} + f \frac{\partial\varphi}{\partial f} = 0, \quad X_2\varphi = aK \frac{\partial\varphi}{\partial K} + bL \frac{\partial\varphi}{\partial L} + f \frac{\partial\varphi}{\partial f} = 0. \quad (6)$$

Clearly, the vector fields X_1, X_2 form a two-dimensional integrable distribution on \mathbb{R}_+^2 : $[X_1, X_2] = \rho_1 X_1 + \rho_2 X_2$, where $\rho_1 = \rho_2 = 0$. The corresponding total differential equation is given by (see Chapter VII, Sato [5] for more details)

$$(fL - bfL)dK + (afK - fK)dL + (bKL - aKL)df = 0,$$

or,

$$(1 - b)\frac{dK}{K} + (a - 1)\frac{dL}{L} + (b - a)\frac{df}{f} = 0. \tag{7}$$

Integrating (7), we arrive at a Cobb-Douglas function of the form (2), where the elasticity coefficients

$$\alpha = \frac{1 - b}{a - b}, \quad \beta = \frac{a - 1}{a - b}$$

satisfy the condition of constant returns to scale $\alpha + \beta = 1$. Of course, one has to also assume that the parameters of the exponential growth a and b are such that the coefficients of elasticity $\alpha, \beta > 0$.

Unfortunately, in spite of much ingenuity employed and a positive result, Sato’s approach based on analytical methods cannot be merged with the approach by Cobb and Douglas based on a data analysis method. Indeed, the data presented in Table 1 represents growth only in one sector of an economy and as such is incompatible with any approach based on the notion of the simultaneous holothenticity. At the same time, it is obvious that an additional equation must be employed to derive the Cobb-Douglas aggregate production function with economically meaningful elasticity coefficients α and β in (2). To resolve this contradiction, the authors of this article employed the bi-Hamiltonian approach in [8] to build on the approach introduced by Sato to derive the Cobb-Douglas production function by analytic methods.

The following is a brief review of the derivation of the Cobb-Douglas production function performed in [8]. Indeed, let us begin with Sato’s assumption about exponential growth in production, labor and capital and rewrite the PDE (4) as the following system of ODEs:

$$\dot{x}_i = b_i x_i, \quad i = 1, 2, 3, \tag{8}$$

where $x_1 = L$ (labor), $x_2 = K$ (capital), $x_3 = f$ (production), $b_1 = b, b_2 = a$ and $b_3 = 1$ in Sato’s notations (see (4)). Next, we rewrite (8) as the following Hamiltonian system:

$$\dot{x}_i = X_H^i = \pi_1^{i\ell} \frac{\partial H}{\partial x_\ell}, \quad i = 1, 2, 3. \tag{9}$$

Here

$$\pi = -x_i x_j \frac{\partial}{\partial x_i} \wedge \frac{\partial}{\partial x_j}, \quad i, j = 1, 2, 3 \tag{10}$$

is the quadratic (degenerate) Poisson bi-vector that defines the Hamiltonian function

$$H = \sum_{k=1}^3 c_k \ln x_k \tag{11}$$

via $X_H = [\pi, H]$, in which the parameters c_k are solutions to the rank 2 algebraic system $A\mathbf{c} = \mathbf{b}$ determined by the skew-symmetric 3×3 matrix A

$$A = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix},$$

$\mathbf{c} = [c_1, c_2, c_3]^T$ with all $c_k > 0$, and $\mathbf{b} = [b_1, b_2, b_3]^T$, satisfying the condition

$$b_1 + b_3 = b_2. \tag{12}$$

Alternatively, we can introduce the following new variables

$$v_i = \ln x_i, \quad i = 1, 2, 3, \tag{13}$$

which lead to an even simpler form of the system (8), namely

$$\dot{v}_i = b_i, \quad i = 1, 2, 3. \tag{14}$$

Interestingly, the substitution (13) is exactly the one used by Cobb and Douglas in [1]. Note that (14) is also a Hamiltonian system, provided $b_1 + b_3 = b_2$, defined by the corresponding (degenerate) Poisson bi-vector $\tilde{\pi}$ with components

$$\tilde{\pi}^{ij} = -\frac{\partial}{\partial v_i} \wedge \frac{\partial}{\partial v_j}$$

and the corresponding Hamiltonian

$$\tilde{H} = \sum_{k=1}^3 c_k v_k.$$

Observing that the function H given by (11) is a constant of the motion of the Hamiltonian system (9), and then solving the equation $\sum_{k=1}^3 c_k \ln x_k = H = \text{const}$ for x_3 , we arrive at the Cobb-Douglas production function (2) after the identification $x_1 = L$, $x_2 = K$, $x_3 = f$, $A = \exp\left(\frac{H_1}{c_3}\right)$, $\alpha = -\frac{c_1}{c_3}$, $\beta = -\frac{c_2}{c_3}$. Next, introduce the following bi-Hamiltonian structure for the dynamical system (8):

$$\dot{x}_i = X_{H_1, H_2} = [\pi_1, H_1] = [\pi_2, H_2], \quad i = 1, 2, 3, \tag{15}$$

where the Hamiltonian functions H_1 and H_2 given by

$$H_1 = b \ln x_1 + \ln x_2 + a \ln x_3, \quad H_2 = \ln x_1 + a \ln x_2 + b \ln x_3. \quad (16)$$

correspond to the Poisson bi-vectors π_1 and π_2

$$\pi_1 = a_{ij}x_i x_j \frac{\partial}{\partial x_i} \wedge \frac{\partial}{\partial x_j}, \quad \pi_2 = b_{ij}x_i x_j \frac{\partial}{\partial x_i} \wedge \frac{\partial}{\partial x_j}, \quad i, j = 1, 2, 3 \quad (17)$$

respectively under the conditions

$$\begin{cases} bb_1 + b_2 + ab_3 = 0, \\ b_1 + ab_2 + b_3b = 0. \end{cases} \quad (18)$$

Note the conditions (18) (compare them to (12)) assure that π_1 and π_2 are indeed Poisson bi-vectors compatible with the dynamics of (8) and corresponding to the Hamiltonians H_1 and H_2 given by (16) respectively. Solving the linear system (18) for a and b under the additional condition $b_1b_2 - b_3^2 \neq 0$, we arrive at

$$a = \frac{b_2b_3 - b_1^2}{b_1b_2 - b_3^2}, \quad b = \frac{b_1b_3 - b_2^2}{b_1b_2 - b_3^2}. \quad (19)$$

Consider now the first integral H_3 given by

$$H_3 = H_1 - H_2 = (b - 1) \ln x_1 + (1 - a) \ln x_2 + (a - b) \ln x_3. \quad (20)$$

Solving the equation $H_3 = \text{const}$ determined by (20) for x_3 , we arrive at the Cobb-Douglas function (2) with the elasticities of substitution α and β given by

$$\alpha = \frac{a - 1}{a - b}, \quad \beta = \frac{1 - b}{a - b}, \quad (21)$$

where a and b are given by (19). Note $\alpha + \beta = 1$, as expected. Also, $\alpha, \beta > 0$ under the additional condition $b_2 > b_3 > b_1$, which implies by (8) that capital ($x_2 = K$) grows faster than production ($x_3 = f$), which, in turn, grows faster than labor ($x_1 = L$). We have also determined the corresponding formula for total factor productivity A (27) - see below for the numerical value of A .

In what follows we will show that the formulas obtained above via the bi-Hamiltonian approach can in fact be matched with the data employed by Cobb and Douglas in [1].

3 Main Result

Solving the separable dynamical system (8), we obtain

$$x_i = c_i \exp(b_i t), \quad i = 1, 2, 3, \quad (22)$$

where $c_i \in \mathbb{R}_+$ and b_i we will determine from the data presented in Table 1.

Taking the logarithm (actually, much like Cobb and Douglas treated their data in [1]!) of both sides in each equations, we linearize them as follows:

$$\ln x_i = C_i + b_i t, \quad i = 1, 2, 3, \quad (23)$$

where $C_i = \ln c_i$.

Our next goal is to recover the corresponding values of the coefficients C_i , b_i , $i = 1, 2, 3$ from the data presented in Table 1. Employing R (see Appendix for more details) and the method of least squares, we arrive at the following values:

$$\begin{aligned} b_1 &= 0.025496, \quad C_1 = 4.669533 \text{ (labor)}, \\ b_2 &= 0.064725, \quad C_2 = 4.612136 \text{ (capital)}, \\ b_3 &= 0.035926, \quad C_3 = 4.664153 \text{ (production)}. \end{aligned} \quad (24)$$

We see that the errors, represented by the \$values in Figs. 1a, 1b and 2a, are all less than 1, which suggests that the formulas (23) fit quite well to the data in Table 1. To measure the goodness of fit, consider, for example, the data presented in the second column of Table 1 (capital). The graph relating observed capital versus estimated capital is the subject of Fig. 3a. Employing R, we have verified that the linear regression shows the adjusted R-squared value of the model is 0.9934, which is very close to 1 (see Fig. 3b).

We also note the values of the estimated coefficients satisfy the inequality $b_2 > b_3 > b_1$, which is in agreement with our algorithm based on the bi-Hamiltonian approach. Identifying $x_1 = L$ and $x_2 = K$ from the data and substituting the values of parameters b_i into the Eq. (19), we obtain

$$a = 4.659322, \quad b = -9.104008, \quad (25)$$

which in turn determine the values of α and β via (21) to be

$$\alpha = 0.265875, \quad \beta = 0.734125. \quad (26)$$

Now we can determine the corresponding value of total factor productivity A from the following formula, obtained by solving the equation $H_3 = \text{const}$ determined by (20),

$$A = \exp\left(\frac{H_3}{a - b}\right), \quad (27)$$

where H_3 is a constant along the flow (22) as a linear combination of the two Hamiltonians H_1 and H_2 given by (16).

Next, using the data from Table 1 and formula (20), we employ R to evaluate H_3 , arriving at the following results: the variance of the resulting distribution of values of H_3 is 0.592229 and the mean of the distribution is 0.136555. By letting $H_3 = 0.136555$ and using (27), the value of A is found to be $A = 1.009971 \approx 1.01$ (compare with (3)).

Therefore, we conclude that using statistical methods we have fitted the differential equations (22) to the values of the elasticities of substitution and total factor productivity obtained via the bi-Hamiltonian approach and the data originally studied by Cobb and Douglas in 1928. In addition, we have demonstrated that Sato’s assumption about exponential growth in production and factors of production [5] is compatible with the results by Cobb and Douglas based on the statistical analysis of the data from the US manufacturing.

Acknowledgements The first author (RGS) wishes to thank the organizers for the invitation to participate in the V AMMCS International Conference and present our results in the sessions “Applied Analysis & Inverse Problems” and “Applications of Dynamical Systems & Differential Equations”, as well as to acknowledge useful discussions pertinent to this research with Professors Herb Kunze (Guelph) and Davide La Torre (Milan). We also wish to thank Professor Ruzyo Sato (NYU) for his interest in our research, comments and suggestions.

Appendix

See Figs. 1, 2a, and 3.

```
> myfun=function(par,data){
+ l = data$labour
+ t = data$year
+ func=sum((l-par[2]-(par[1]*(t-1899)))^2)
+ return(func)
+ }
> optim(myfun, par=c(0.1,4.605170),data=mydata)
$par
[1] 0.02549605 4.66953271

$value
[1] 0.1827943

$counts
function gradient
      59      NA

$convergence
[1] 0

$message
NULL
```

(a) Labor fitting.

```
> myfun=function(par,data){
+ k = data$capital
+ t = data$year
+ func=sum((k-par[2]-(par[1]*(t-1899)))^2)
+ return(func)
+ }
> optim(myfun, par=c(0.1,4.605170),data=mydata)
$par
[1] 0.06472564 4.61213569

$value
[1] 0.03065574

$counts
function gradient
      65      NA

$convergence
[1] 0

$message
NULL
```

(b) Capital fitting.

Fig. 1 Labor and capital

```

> myfun=function(par, data) {
+ p = data$output
+ t = data$year
+ func=sum((p-par[2]-(par[1]*(t-1899)))^2)
+ return(func)
+ }
> optim(myfun, par=c(0.1,4.605170),data=mydata)
$par
[1] 0.03592651 4.66415344

$value
[1] 0.1825852

$counts
function gradient
63 NA

$convergence
[1] 0

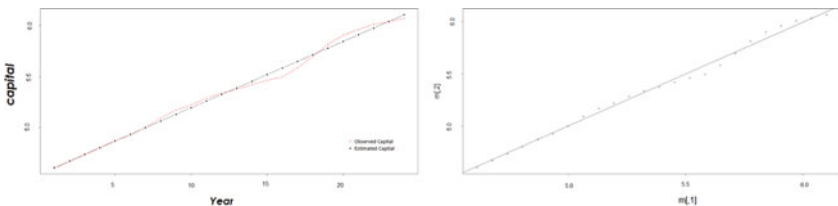
$message
NULL

> predictY = function(par, data){
+ k = data$capital
+ l = data$labour
+ p = data$output
+ func = (par[2]-1)*1+(1-par[1])*k+(par[1]-par[2])*p
+ return(func)
+ }
> m=predictY(par=c(4.659322,-9.104008), data= mydata)
> plot(m)
> mean(m)
[1] 0.1365547
    
```

(a) Production fitting.

(b) Total factor productivity fitting.

Fig. 2 Production and total factor productivity



(a) Observed capital versus estimated capital.

(b) The linear regression of the observed and estimated capital from 1899 to 1922.

Fig. 3 Linear regression

References

1. Cobb, C.W., Douglas, P.H.: A theory of production. *American Econ. Rev.* **8**, 139–1965 (1928)
2. Felipe, J., Adams, F.G.: The estimation of the Cobb-Douglas function: A retrospective review. *Eastern Econ. J.* **31**, 427–445 (2005)
3. Douglas, P.H.: The Cobb-Douglas production function once again: Its history, its testing, and some new empirical values. *J. Polit. Econ.* **84**, 903–915 (1976)
4. Fitting of Cobb-Douglas production functions: Prajneshu. Revisited. *Agric. Econ. Res. Rev.* **21**, 289–292 (2008)
5. Sato, R.: *Theory of Technical Change and Economic Invariance*. Academic Press, New York (1981)
6. Sato, R., Ramachandran, R.V.: *Symmetry and Economic Invariance*, 2nd edn. Springer, New York (2014)
7. Smirnov, R., Wang, K.: In search of a new economic model determined by logistic growth. *European J. Appl. Math.* (2019). <https://doi.org/10.1017/S0956792519000081>
8. Smirnov, R., Wang K.: The Hamiltonian approach to the problem of derivation of production functions in economic growth theory. (2019) arxiv.org/abs/1906.11224

Inferring Rankings from First Order Marginals



Sarah Wolff

Abstract Motivated by applications in ranked-choice voting, we consider the problem of recovery of an election profile—encoded by a function f on the symmetric group—given only partial data. In particular, we investigate the combinatorial structure of the matrix of first order marginals, which gives the number of votes cast that ranked each alternative in each position. We investigate conditions on f that allow us to exploit this combinatorial structure to recover the original function f . As the matrix of first order marginals is the Fourier coefficient of the permutation representation of the symmetric group, this work sits within the context of algebraic compressed sensing, which tackles the question of how to recover a sparse function f on a finite group given only a subset of the Fourier coefficients of f .

Keywords Ranked data · Discrete fourier transform · Symmetric group

1 Introduction

Consider an election procedure in which k members of a society select a ranking amongst n alternatives. Let S_n denote the set of all possible rankings of the n alternatives. The set of rankings given by the individual members is called a *profile* and can be encoded by a function $f : S_n \rightarrow \mathbb{Z}^+$, giving the number, $f(\sigma)$, of votes cast for each ranking σ . Suppose, however, that the only information known is the number of votes cast that ranked each alternative in each position. Can one uniquely recover the original profile f ? Thinking of f as a probability distribution on S_n (with appropriate normalizing), this becomes the question of recovery of f given its matrix $Q1$ of first order marginals, where $Q1_{ij}$ gives the number of members who ranked alternative j in position i .

Taking an algebraic point of view, we recognize the set S_n , along with the operation of composition, as the symmetric group. The representation theory of the symmetric

S. Wolff (✉)

Denison University, 100 W. College Street, Granville, OH 43023, USA
e-mail: wolffs@denison.edu

© Springer Nature Switzerland AG 2021

D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343, https://doi.org/10.1007/978-3-030-63591-6_67

735

group is well-understood and has been applied fruitfully in social choice theory over the years. For example, Kelly [12] rephrases the arrowian framework using the symmetric group; more recently, Bubboloni and Gori [2, 3] have used subgroup actions to determine when election procedures satisfy such properties as reversal-symmetry, anonymity, and neutrality.

A line of work taken up independently by Diaconis, Orrison, and Saari [6, 7, 14], among others, considers the *Fourier analysis* of functions corresponding to election profiles. Analyzing a function on S_n by considering its *Fourier coefficients*, i.e. its projection into submodules determined by the irreducible representations of S_n , yields new information about the function. For example, in [7] Diaconis uses Fourier analysis on the 1980 American Psychological Association (APA) presidential election data, finding a large effect coming from two pairs of alternatives. The APA at the time was divided primarily among academicians and clinicians; Diaconis's results show that voters primarily lined up behind either the academician pair or the clinician pair.

In some sense, our work takes the opposite perspective: in Sect. 2 we see that the matrix $Q1$ is exactly the Fourier coefficient $\hat{f}(\rho)$ for ρ the permutation representation of S_n . The motivating question then becomes: given only a subset of Fourier coefficients of f , can one uniquely recover f ?

The same question, in the context of the cyclic group C_N rather than the symmetric group S_n , forms the heart of *compressed sensing*, which originated from the question of reconstructing a discrete time signal from a small number of frequencies, i.e., a subset of its Fourier coefficients. Introduced by Donoho [8] and in a series of papers by Candés, Romberg, and Tao (including [4, 5]), the central methods of compressed sensing are not only of theoretical interest but have also been used in applied settings such as signal processing, acoustic imaging, and medical imaging, among others [1].

Given only a subset of Fourier coefficients there are often infinitely many functions that share these Fourier coefficients. Candés, Romberg, and Tao proposed seeking the solution with “minimum complexity” whose partial Fourier coefficients match those of f . They show that the function with minimum ℓ_1 norm *exactly* recovers f with overwhelming probability when $|\text{supp}(f)| \leq |S'|/\log N^6$, for $|S'|$ the number of Fourier coefficients (see p. 1441 of [4] and [5]).

In the social choice context of a function f on the symmetric group S_n , Jagathula and Shah [9] connect to the area of compressed sensing by focusing on finding the sparsest solution consistent with the given Fourier coefficients. They derive two conditions, *linear independence* and *unique witness*, which are sufficient for functions to be recovered exactly from the matrix $Q1$ and they provide an algorithm to recover such functions. Jagathula and Shah show that functions with sparsity K are recoverable with high probability for $K \leq (1 - \epsilon)n \log n$. This result indicates that the conditions are indeed not very restrictive for functions with sparsity K , paralleling similar results in the classical (C_N) case.

While these results parallel those in the classical compressed sensing literature, in social choice applications it is unlikely for a function f corresponding to a profile to

satisfy these sparsity bounds, as this implies that the votes are concentrated at fewer than $(1 - \epsilon)n \log n$ rankings. Especially in the case when f has nonnegative integer outputs and bounded domain (as in the context of elections), the likelihood that f satisfies the linear independence property decreases significantly as the size of the support increases.

We remove the linear independence property of [9] as a starting point to consider the recovery of more general functions on S_n given a subset of Fourier coefficients, focusing here on the matrix $Q1$ described above. The techniques of [9] no longer apply in this setting as their algorithm leans heavily on the linear independence property (see Sect. 3.1). We instead find and characterize combinatorial structure within the matrices, allowing for the recovery of the original function f by classifying patterns within the matrix. We also use the combinatorial structure of the matrices to show that when unique recovery of f is impossible, the class of functions sharing the Fourier coefficient $Q1$ still share properties that inform the outcome of the associated election. In Sect. 2 we introduce the necessary definitions and background material. In Sect. 3.1 we describe the conditions on our functions and derive the combinatorial structure of the matrices. In Sect. 3.2 we investigate the question of uniqueness, asking if and when functions satisfying the conditions of Sect. 3.1 share a $Q1$ matrix. While we restrict ourselves to work in a specialized setting due to space considerations, the combinatorial results and structure we derive extend to the more general setting. We describe the connection in Sect. 4 and conclude with further directions and questions. While we give sketches of most proofs, for brevity's sake we omit the full details.

2 Background and the Fourier Transform

Consider an election procedure in which members of a society select a ranking amongst n alternatives. Let S_n denote the set of all possible rankings of the n alternatives. The set of rankings given by the individual members is called a *profile* and can be encoded by a function $f : S_n \rightarrow \mathbb{Z}^+$, giving the number, $f(\sigma)$, of votes cast for each ranking σ .

In what follows, we will use one-line notation to denote a ranking; e.g., $f(1324) = 7$ indicates that 7 people ranked alternative 1 first, alternative 2 third, alternative 3 s, and alternative 4 fourth.

Definition 1 Let $f : S_n \rightarrow \mathbb{Z}^+$. The *Q1 matrix of f* , denoted Q^f , is the matrix whose (i, j) th entry is given by $Q_{ij}^f = \sum_{\sigma:\sigma(j)=i} f(\sigma)$.

In other words, the ij th entry of Q^f gives the number of members who ranked alternative j in position i .

Example 1 Let $f : S_5 \rightarrow \mathbb{Z}^+$ with nonzero function values given below.

12435	25431	52431	31425	43125	34125	45231	45321	42531	32415	32145	32451	45132	45123	35124
1	9	10	2	3	7	8	4	11	15	19	5	13	41	47

$$\text{Then } Q^f = \begin{pmatrix} 1 & 2 & 130 & 15 & 47 \\ 9 & 61 & 8 & 104 & 13 \\ 95 & 3 & 4 & 52 & 41 \\ 80 & 7 & 42 & 19 & 47 \\ 10 & 122 & 11 & 5 & 47 \end{pmatrix}.$$

We will use the function of Example 1 as our running example.

While it is natural to think about f as taking positive integer values, realizing that S_n forms a mathematical group, the *symmetric group*, and that f is an element of the *group algebra* $\mathbb{C}[S_n]$ of complex-valued functions on S_n gives the important connection to Fourier analysis on finite groups. While we make the following definitions for arbitrary finite groups, S_n will be the group of focus in this paper.

Definition 2 Let G be a finite group. A *matrix representation* ρ of G is a function from G to the set of $n \times n$ invertible matrices such that $\rho(ab) = \rho(a)\rho(b)$ for all elements a, b of G . We say that n is the *dimension* of ρ .

Definition 3 Let G be a finite group and $f \in \mathbb{C}[G]$. Let ρ be a matrix representation of G . The **Fourier transform of f at ρ** , denoted $\hat{f}(\rho)$, is the matrix sum

$$\hat{f}(\rho) = \sum_{s \in G} f(s)\rho(s).$$

For $G = C_N$ the cyclic group of order N , all irreducible representations are 1-dimensional and the set of Fourier transforms corresponding to a complete set of inequivalent irreducible representations of C_N is the usual discrete Fourier transform. For $G = \{0, 1\}^n$ the Boolean Cube, Fourier analysis of functions on G has had many interesting applications to social choice theory (see, for example [10, 11, 13].

Let $G = S_n$ and define the representation ρ as follows: for $\sigma \in S_n$, $\rho(\sigma)_{ij} = 1$ if $\sigma(j) = i$, and 0 otherwise. This is the *permutation representation* of S_n . While not irreducible, ρ decomposes into the 1-dimensional trivial representation and the $(n - 1)$ -dimensional standard representation: $\rho = \tau_{(n)} \oplus \tau_{(n-1,1)}$.

Note that for ρ the permutation representation of S_n and f a function on S_n , the $Q1$ matrix of f is given by $\sum_{\sigma \in S_n} f(\sigma)\rho(\sigma)$. In other words, the $Q1$ matrix of f is the Fourier coefficient $\hat{f}(\rho)$ of f at the representation ρ . The central question of this work is then how to recover f given the Fourier coefficient of f at the permutation representation.

3 Main Results

3.1 Structure of the Matrices

Given a function $f : S_n \rightarrow \mathbb{C}$, the support of f , denoted $\text{supp}(f)$, is the set of inputs σ such that $f(\sigma) \neq 0$, i.e., the set of rankings chosen by at least one member. We impose the following conditions on f :

1. $f(\sigma) \neq f(\tau)$ for any $\sigma \neq \tau \in \text{supp}(f)$.
2. For each $\sigma \in \text{supp}(f)$, there exists $i, j \in \{1, \dots, n\}$ such that $\sigma(i) = j$ while $\tau(i) \neq j$ for all $\tau \in \text{supp}(f)$, $\tau \neq \sigma$. We will call such a pair (i, j) a **unique witness** for σ .

Condition 2 is the *unique witness* property of [9] while Condition 1 is weaker than, but implied by, the *linear independence* property of [9]. Note that a permutation $\sigma \in \text{supp}(f)$ could have multiple unique witnesses.

Definition 4 Let $f : S_n \rightarrow \mathbb{C}$ and let Q^f be the $Q1$ matrix of f . Let $\sigma \in \text{supp}(f)$. An entry Q^f_{ij} of Q^f is a **unique witness function value** of Q^f (corresponding to σ) if $Q^f_{ij} = f(\sigma)$ and $\sigma(j) = i$. All other nonzero entries of Q^f are **nonfunction values**. A collection of unique witness function values $\{Q^f_{i_1j_1}, \dots, Q^f_{i_kj_k}\}$ corresponding to distinct $\sigma_1, \dots, \sigma_k \in \text{supp}(f)$ is a set of **distinct unique witness function values** of Q^f .

Weakening the linear independence property to Condition 1 removes two properties of Q^f that are crucial to the algorithm of [9]. First, if $Q^f_{ij} = f(\sigma)$ for some $\sigma \in \text{supp}(f)$, this does not necessarily imply $\sigma(j) = i$. Moreover, if for some $I \subset [n]$, $Q^f_{ij} = \sum_{\sigma \in I} f(\sigma)$, then $\sigma_k(j) = i$ is not guaranteed for all $k \in I$.

Example 2 Given the matrix Q^f of Example 1, the knowledge that Q^f_{45} is a unique witness function value for Q^f immediately implies that there exists $\sigma \in \text{supp}(f)$ with $\sigma(5) = 4$ and $f(\sigma) = 47$. However, each entry of 47 in Q^f does not necessarily correspond to an edge of σ ; indeed, $Q^f_{15} = Q^f_{55} = 47$ but this cannot imply $\sigma(5) = 1$ or that $\sigma(5) = 5$. Moreover, for $\sigma_1 = 31425$, $\sigma_2 = 42531$, $\sigma_3 = 32415$, and $\sigma_4 = 32145$, $Q_{15} = \sum_{i=1}^4 f(\sigma_i)$, but this does not imply $\sigma_i(5) = 1$ for $1 \leq i \leq 4$.

We restrict our attention to functions that have maximum support, as this setting informs the less restrictive scenarios. In Sect.4 we describe how to extend these results to the general setting. We also focus on nonnegative outputs as in [9].

Lemma 1 Let $n > 3$ and let $k_n = \max_f |\text{supp}(f)|$ over all $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfying Conditions 1 and 2. Then $k_n = n(n - 2)$.

Proof (Sketch of proof). Let $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfy Conditions 1 and 2 and let Q^f be the $Q1$ matrix of f . If Q^f has n distinct unique witness function values in a single column, this immediately implies $|\text{supp}(f)| = n$. If Q^f has $(n - 1)$ distinct unique witness function values in at least one column, this forces $|\text{supp}(f)| < n(n - 2)$ for

$n > 3$. If Q^f has fewer than $(n - 2)$ distinct unique witness function values in a column, f cannot have maximum support because $|\text{supp}(f)| \geq n(n - 2)$ only if a column of Q^f has greater than $(n - 2)$ distinct witness function values, placing it in the first two scenarios. Thus, if Q^f has exactly $(n - 2)$ distinct unique witness function values in each column, $|\text{supp}(f)|$ is maximized (see Example 3).

Example 3 As we describe below, the support of $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ is completely determined by the placement of circles that represent a distinct unique witness function value of Q^f . Any choice of a distinct positive real number for each $\sigma \in \text{supp}(f)$ gives a function that satisfies Conditions 1, 2 and has support $n(n - 2)$.

$$\begin{pmatrix} \square & \bigcirc & \bigcirc & \bigcirc & \dots & \bigcirc & \square \\ \square & \square & \bigcirc & \bigcirc & \dots & \bigcirc & \bigcirc \\ \bigcirc & \square & \square & \bigcirc & \dots & \bigcirc & \bigcirc \\ \bigcirc & \bigcirc & \square & \square & \dots & \bigcirc & \bigcirc \\ & & & \vdots & & & \\ \bigcirc & \bigcirc & \bigcirc & \bigcirc & \dots & \square & \bigcirc \\ \bigcirc & \bigcirc & \bigcirc & \bigcirc & \dots & \square & \square \end{pmatrix}$$

Suppose Q^f_{ij} is circled. Then $Q^f_{ij} = f(\sigma)$ for $\sigma \in \text{supp}(f)$ with $\sigma(j) = i$. As Q_{jj} is a nonfunction value and $\sigma(j) \neq j$, this implies $\sigma(j - 1) = j$ because the only other nonfunction value in row j is in column $j - 1$. Continuing in this manner, if $i > j, \sigma(k) = (k + 1) \bmod n$ for $1 \leq k \leq j - 1$ and for $i \leq k \leq n$. Else, $\sigma(k) = k$. If $i < j, \sigma(k) = k$ for $1 \leq k \leq i - 1$ and for $j + 1 \leq k \leq n$. Else, $\sigma(k) = k + 1$.

For $n > 3$ let $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfy Conditions 1 and 2 with maximum support. By Lemma 1, there exists a set of exactly $n(n - 2)$ distinct unique witness function values of Q^f , and the proof shows each column (respectively, row) of Q^f contains exactly $(n - 2)$ of them, while the rest are nonfunction values. We show that each nonfunction value Q^f_{ij} can be written as $Q^f_{ij} = \sum_{k \in I} f(\sigma_k)$, for I the **(n-2)-triangle** of distinct unique witness function values corresponding to Q^f_{ij} .

Definition 5 Let Q be an $n \times n$ matrix. An **(n-2)-triangle of Q** is a choice of two rows and two columns, along with a numbering of the remaining rows R_1, \dots, R_{n-2} , and columns C_1, \dots, C_{n-2} and a set of matrix values T in these numbered rows and columns, such that each column C_i and row R_i contains exactly i of the entries of T .

Example 4 For our running example, choose row 1, row 2, column 1, column 3, number the remaining rows and columns as follows, and let $T = \{2, 5, 7, 15, 19, 47\}$.

$$\begin{matrix} & & C_2 & & C_3 & C_1 \\ R2 & \begin{pmatrix} 1 & \textcircled{2} & 130 & \textcircled{15} & 47 \\ 9 & 61 & 8 & 104 & 13 \\ 95 & 3 & 4 & 52 & 41 \end{pmatrix} \\ R3 & \begin{pmatrix} 80 & \textcircled{7} & 42 & \textcircled{19} & \textcircled{47} \end{pmatrix} \\ R1 & \begin{pmatrix} 10 & 122 & 11 & \textcircled{5} & 47 \end{pmatrix} \end{matrix}$$

Note 1 The subset T of matrix entries completely determines the omitted rows and columns along with the numbering of the remaining rows and columns. For this reason we will often refer to a $(n - 2)$ -triangle by simply identifying T .

For Q an $n \times n$ matrix let $[Q]_{i,j}$ denote the submatrix of Q obtained by deleting the i th row and j th column of Q . We will use bold font to represent the new index of a row or column of Q in $[Q]_{i,j}$. In other words, column k (respectively, row k) of Q^f becomes column \mathbf{k} (respectively, row \mathbf{k}) of $[Q]_{i,j}$, where $\mathbf{k} = k - 1$ if $k > j$ (respectively, $k > i$), and otherwise $\mathbf{k} = k$.

Theorem 1 *Let $n > 3$ and let $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ be a function satisfying Conditions 1 and 2 with maximum support. Let Q_{ij}^f be a nonfunction value of Q^f . Then there is exactly one $(n - 2)$ -triangle of Q^f whose corresponding set T is comprised of distinct unique witness function values of Q^f that occur in the submatrix $[Q^f]_{i,j}$. Moreover $\sigma \in \text{supp}(f)$ has $\sigma(j) = i$ if and only if $f(\sigma) \in T$.*

Example 5 Before sketching the proof of Theorem 1 we provide an example. Consider the function of Example 1 and consider the nonfunction value $Q_{31}^f = 95$. Note that 6 permutations $\sigma \in \text{supp}(f)$ have $\sigma(1) = 3$, and that the function values $f(\sigma)$ for these permutations are 2, 5, 7, 15, 19, 47. We see immediately that these are the distinct unique witness function values circled in Example 4 and that Q_{31}^f is the sum of these values.

Proof (Sketch of proof). The proof proceeds by induction. We will focus on the induction step in this sketch. Let $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ be a function satisfying the conditions of Theorem 1. Let Q^f be the $Q1$ matrix of f and let Q_{ij}^f be a nonfunction value of Q^f . Let Q_{ik}^f be the additional nonfunction value of Q^f in row i , let $Q_{\ell k}^f$ be the additional nonfunction value in column k , and let Q_{mj}^f be the additional nonfunction value in column j (see Example 6). It can be shown that $Q_{\ell j}^f$ is a distinct unique witness function value of Q^f ; essentially, no two rows or columns can share the same nonfunction values.

Consider the submatrix $[Q^f]_{i,k}$. This is an $(n - 1) \times (n - 1)$ matrix with exactly two nonfunction values in each row except for row ℓ and in each column except column \mathbf{j} . We define a function $g : S_{n-1} \rightarrow \mathbb{R}_{\geq 0}$ whose $Q1$ matrix has nonfunction values in the same locations as those of f in $[Q^f]_{i,k}$, along with one additional nonfunction value in row ℓ , column \mathbf{j} .

In essence, we restrict each $\sigma \in \text{supp}(f)$ with unique witness not (\cdot, i) , (k, \cdot) , or (j, ℓ) to $\sigma|_{n-1} \in S_{n-1}$ by restricting σ to the submatrix $[Q^f]_{i,k}$: removing column k requires some inputs to be shifted, while removing row i requires some outputs to be shifted and also requires us to send $\sigma^{-1}(i)$ elsewhere (to ℓ).

We then define $g : S_{n-1} \rightarrow \mathbb{R}_{\geq 0}$ so that $\rho \in \text{supp}(g)$ only if $\rho = \sigma|_{n-1}$ for $\sigma \in \text{supp}(f)$ with unique witness not (\cdot, i) , (k, \cdot) , or (j, ℓ) , and in this case we let $g(\rho) = f(\sigma)$. See Example 6 for an example of g .

We show that g is well-defined, satisfies Conditions 1 and 2, and $|\text{supp}(g)| = (n - 1)(n - 3)$, so by induction the statement of Theorem 1 holds for g . Then for the

nonfunction value $Q_{\ell j}^g$ in the $Q1$ matrix for g , there exists exactly one occurrence of an $(n - 3)$ -triangle in $[Q^g]_{\ell, j}$ whose corresponding set T_1 contains only distinct unique witness function values of Q^g (in Example 6, $T_1 = \{2, 7, 47\}$). In other words, T_1 contains all function values $g(\sigma|_{n-1})$ such that $\sigma|_{n-1}(\mathbf{j}) = \ell$. By definition of $\sigma|_{n-1}$, this implies that for each $\sigma \in \text{supp}(f)$ with unique witness not (j, ℓ) , (\cdot, i) , or (k, \cdot) , $f(\sigma) \in T_1$ if and only if $\sigma(j) = i$.

Letting T_2 be the set of distinct unique witness function values in column k of Q^f , (in Example 6, $T_2 = \{5, 15, 19\}$), we then show that $T = T_1 \cup T_2$ forms an $(n - 2)$ -triangle of Q^f .

Example 6 Consider the function of Example 1 and let $i = 3, j = 1$, with $Q_{31}^f = 95$. Then $k = 4, \ell = 2, m = 4, Q_{ik}^f = 52, Q_{\ell k}^f = 104, Q_{mj}^f = 80$, and $[Q^f]_{i, k}$ is below. Note that $Q_{ij}^f = 9$ and $\ell = 2, \mathbf{j} = 1$, so the distinct unique witness function value in the second row and first column of $[Q^f]_{i, k}$ is 9.

For $\sigma = 12435, \sigma|_4 = 1234$. Since $f(\sigma) = 1, g(1234) = 1$. Similarly, $g(2134) = 2, g(3421) = 8, g(3412) = 13, g(2314) = 7, g(2413) = 47, g(4231) = 10, g(3241) = 11$. The $Q1$ matrix for g is below, with nonfunction values in the same location as $[Q^f]_{i, k}$ along with additional nonfunction value $Q_{\ell j}^g$. The $(n - 3)$ triangle for $Q_{ij} = 56 (T_1)$ is in red.

$$[Q^f]_{i, k} = \begin{bmatrix} \textcircled{1} & \textcircled{2} & 130 & 47 \\ \textcircled{9} & 61 & \textcircled{8} & \textcircled{13} \\ 80 & \textcircled{7} & 42 & \textcircled{47} \\ \textcircled{10} & 122 & \textcircled{11} & 47 \end{bmatrix} \quad Q^g = \begin{bmatrix} \textcircled{1} & \textcircled{2} & 67 & 29 \\ 56 & 22 & \textcircled{8} & \textcircled{13} \\ 32 & \textcircled{7} & 13 & \textcircled{47} \\ \textcircled{10} & 68 & \textcircled{11} & 10 \end{bmatrix}$$

Corollary 1 Let $n > 3$ and let $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ be a function satisfying Conditions 1 and 2 with maximum support. Given Q^f and the locations of the distinct unique witness function values of Q^f , f is completely recoverable.

Example 7 For f the function of Example 1, let $\sigma \in \text{supp}(f)$ with $f(\sigma) = 2$. In Example 6 and the proof sketch of Theorem 1 we saw that nonfunction value $Q_{31}^f = 95$ has 3-triangle $T = \{2, 5, 7, 15, 19, 47\}$. We could similarly find that the non-function values $Q_{43}^f = 42$ and $Q_{55}^f = 47$ have 3-triangles $T' = \{1, 2, 5, 9, 10, 15\}$ and $T'' = \{1, 2, 3, 7, 15, 19\}$, respectively. These 3-triangles give 3 edges for the permutations of their shared function values; in particular, since $2 \in T \cap T' \cap T''$, $\sigma(1) = 3, \sigma(3) = 4$, and $\sigma(5) = 5$. The distinct unique witness function value of $\sigma (Q_{12}^f)$ is then enough to completely determine $\sigma = 31425$, which matches the original function f of Example 1. Continuing this process recovers the remaining permutations.

3.2 ‘Uniqueness’

One might ask how many functions satisfying Conditions 1 and 2 with maximum support can have the same $Q1$ matrix. Somewhat surprisingly, a maximum of two such functions can share a $Q1$ matrix; additionally, the two functions share properties of interest in social choice theory (see Sect. 4).

Definition 6 A **circling** of an $n \times n$ matrix Q is a choice of $n(n - 2)$ matrix entries so that there exists a function $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfying Properties 1 and 2 with $Q^f = Q$ and a complete set of distinct unique witness function values of Q^f is chosen.

Note that by Corollary 1 a circling of an $n \times n$ $Q1$ matrix completely determines the corresponding function f .

Example 8 The following two circlings of Q yield distinct functions f and g with $Q^f = Q = Q^g$. We use Corollary 1 to recover f and g , given below.

$$\left(\begin{array}{cccc} \textcircled{5} & \textcircled{6} & 18 & 19 \\ 23 & \textcircled{3} & 21 & \textcircled{1} \\ 10 & 22 & \textcircled{2} & \textcircled{4} \\ \textcircled{10} & 17 & \textcircled{7} & 14 \end{array} \right) \quad \left(\begin{array}{cccc} \textcircled{5} & \textcircled{6} & 18 & 19 \\ 23 & \textcircled{3} & 21 & \textcircled{1} \\ \textcircled{10} & 22 & \textcircled{2} & 14 \\ 10 & 17 & \textcircled{7} & \textcircled{4} \end{array} \right)$$

$Q^f \qquad \qquad \qquad Q^g$

σ	3412	2431	3214	1324	3124	2341	4321	2413	σ	4312	2431	4213	1423	4123	2341	3421	2314
$f(\sigma)$	1	2	3	5	6	7	10	14	$g(\sigma)$	1	2	3	5	6	7	10	14

Note 2 A choice of $n(n - 2)$ entries of an $n \times n$ matrix Q so that each row and each column of Q has $n - 2$ circled entries is not necessarily a circling. For example, the proof of Theorem 1 showed that no two columns of Q can be circled identically.

Theorem 2 Let $n > 3$ and let Q be the $Q1$ matrix of a function $f_1 : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfying Properties 1 and 2 with maximum support. Then there is at most one additional function $f_2 : S_n \rightarrow \mathbb{R}_{\geq 0}$ satisfying Properties 1 and 2 with maximum support and $Q1$ matrix Q .

Proof (Sketch of proof). We prove this theorem through a series of lemmas that show an $n \times n$ matrix Q cannot have three distinct circlings, each proof using a similar induction step as in the proof of Theorem 1. The first lemma shows that if an $n \times n$ matrix Q had three distinct circlings, then at least two of the three circlings would differ in more than two columns or more than two rows. The next two lemmas show that this is impossible: two circlings cannot differ in 3 or more columns or 3 or more rows.

4 Generalizations and Further Directions

While we focused here on the case of recovery of a function $f : S_n \rightarrow \mathbb{R}_{\geq 0}$ that satisfies Conditions 1 and 2 with maximum support, the combinatorial structure of the matrices in this setting helps inform the less restrictive settings. We discuss some of our results in these settings below.

For functions with support smaller than $n(n - 2)$, a complete set of distinct unique witness function values has fewer entries in each row and column, so a nonfunction value is no longer guaranteed to come from an $(n - 2)$ -triangle of distinct function values. However, if we consider all unique witness function values, rather than just the distinct ones, we recover results similar to the ones described above. The key difference is we must discard all function values in the same row or column as the nonfunction value. Doing so allows for the identification of function values that contribute to each nonfunction value, again allowing for the recovery of a function with the given $Q1$ matrix.

On the uniqueness side, in the maximum support case, Theorem 2 shows there are at most exactly two functions with the same $Q1$ matrix. Indeed, these functions f and g share some interesting voting theoretic properties. For instance, we show that for any n , for all but three alternatives, an alternative is a *Condorcet winner* in f if and only if she is a Condorcet winner in g . For example, for the functions in Example 8 above, alternative C is a Condorcet winner in both elections. Indeed, choosing any distinct real numbers for the circled entries in this example gives functions f and g where alternative C is either a Condorcet winner in both or in neither. Extending to functions with support smaller than $n(n - 2)$, we have identified families of functions that share a fixed $Q1$ matrix. We are still investigating whether these families similarly share interesting properties such as the Condorcet statement above.

References

1. Boche, H. et al. (eds.): Compressed sensing and its applications. Applied and Numerical Harmonic Analysis. Second International MATHEON Conference 2015. Birkhäuser/Springer, Cham (2017), pp. xix+388
2. Bubboloni, D., Gori, M.: Anonymous and neutral majority rules. *Soc. Choice Welf.* **43**(2), 377–401 (2014)
3. Bubboloni, D., Gori, M.: Symmetric majority rules. *Math. Social Sci.* **76**, 73–86 (2015)
4. Candés, E.J.: Compressive sampling. In: International Congress of Mathematicians, vol. III. Eur. Math. Soc., Zürich, pp. 1433–1452 (2006)
5. Candés, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory* **52**(12), 5406–5425 (2006)
6. Daugherty, Z. et al.: Voting, the symmetric group, and representation theory. *Amer. Math. Monthly* **116**(8), 667–687 (2009)
7. Diaconis, P.: A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17**(3), 949–979 (1989)
8. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inform. Theory* **52**(4), 1289–1306 (2006)
9. Jagabathula, S., Shah, D.: Inferring rankings using constrained sensing. *IEEE Trans. Inform. Theory* **57**(11), 7288–7306 (2011)

10. Gil, K.: Noise sensitivity and chaos in social choice theory. In: Fete of Combinatorics and Computer Science, vol. 20, pp. 173–212 . Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest (2010)
11. Keller, N.: A tight quantitative version of Arrow's impossibility theorem. *J. Eur. Math. Soc. (JEMS)* **14**(5), 1331–1355 (2012)
12. Kelly, J.S.: Symmetry groups. In: *Soc. Choice Welf.* 8.1 (1991), pp. 89–95. *Inferring Rankings From First Order Marginals* 11
13. O'Donnell, R.: Some topics in analysis of Boolean functions. In: *STOC'08*, pp. 569–578 . ACM, New York (2008)
14. Saari, D.G.: Symmetry, voting, and social choice. *Math. Intell.* **10**(3), 32–42 (1988)

High-Frequency Statistical Modelling for Jump-Diffusion Multi-asset Price Processes with a Systemic Component



Rulin Xu and Roman N. Makarov

Abstract This paper is concerned with the statistical modelling of the high-frequency dynamics of financial markets. We study whether a systemic component in a multi-asset price model can explain all jumps in the market's dynamics. By performing statistical analysis of high-frequency data from Wharton Research Data Services (WRDS), we detect disjoint and common jumps in intraday time series of asset prices. We calibrate and compare two multi-asset models with a systemic component: the model with co-jumps only and the model with both common and asset-specific jump components. We assume that jumps follow a compound Poisson process, and the jump sizes have a normal probability distribution. The Kullback–Leibler divergence and the Kolmogorov–Smirnov tests are used to compare the models and show that the model with common and asset-specific jumps provides a better fit to high-frequency data.

Keywords Jump-diffusion process · Multi-asset pricing model · High-frequency trading · Jump and co-jump tests

1 Introduction

With the rapid development of digital technologies, high-frequency data can be accurately collected and analyzed in an efficient manner (see [2, 3, 5]). In [4], a multivariate jump-diffusion was proposed for modelling financial securities with missing or asynchronous data in time series of historical prices. The model was constructed in such a way that low-activity assets correlate with each other only implicitly through the high-activity asset.

R. Xu (✉) · R. N. Makarov
Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: xuxx1870@mylaurier.ca

R. N. Makarov
e-mail: rmakarov@wlu.ca

© Springer Nature Switzerland AG 2021
D. M. Kilgour et al. (eds.), *Recent Developments in Mathematical, Statistical and Computational Sciences*, Springer Proceedings in Mathematics & Statistics 343,
https://doi.org/10.1007/978-3-030-63591-6_68

In this paper, we propose a new multi-asset jump-diffusion model with a systemic component, where each asset has an individual jump component. We apply this multivariate process to model the dynamics of several high-frequently trading assets. The S&P500 stock index can be used to describe the systemic behaviour of the whole financial market. Walmart Inc. (WMT) and Apple Inc. (AAPL) stocks are chosen for our analysis as companies representing different industries.

The Ait-Sahalia and Jacod's jump test [1] is used to find dates for which the intraday price data contain jumps (see also [8, 9]). After that, we employ the Jacod and Todorov's co-jump test [7] to detect common and disjoint jumps when both the asset and the market index have jumps on the same day. After that, we construct the model and compare it with the one from [4]. The performance of each model is analyzed using the Kullback–Leibler divergence and the Kolmogorov–Smirnov test.

2 High-Frequency Data Processing

The S&P500 stock index is commonly considered as an indicator of the U.S. economy as a whole. Two representative high-frequently trading assets with tickets AAPL and WMT are selected for our analysis. The high-frequency intraday price data (at the millisecond level) are obtained from Wharton Research Data Services. All transactions for each trading day from 9:30 am to 4:00 pm in year 2018 were collected.

High-frequency data are irregularly spaced over time and contain market microstructure noise, when the asset price is abnormally rising up and then dropping down (or vice versa). The deviation from fundamental value is a typical characteristic of the high-frequency trading in modern financial markets, which would cause unstable estimates of some parameters. Therefore, we constructed the following procedure to eliminate the microstructure noise. Our approach is similar to the noise mitigating strategy described in [2].

Step 1: Let $S(t)$ denote the asset price at time $t \in \{t_0, t_1, \dots\}$. Let $\Delta X_j = \ln S(t_j) - \ln S(t_{j-1})$ represent the log-return over $[t_{j-1}, t_j]$. Additionally, we standardize the log-returns by subtracting the sample mean $\overline{\Delta X}$ and dividing by the sample standard deviation $s_{\Delta X}$ to obtain $\Delta X_j^* = \frac{\Delta X_j - \overline{\Delta X}}{s_{\Delta X}}$.

Step 2: Choosing 10 as the critical value, we find all indices j_k^* with $k \geq 1$ such that $|\Delta X_{j_k^*}^*| > 10$. Let K denote the number of such indices.

Step 3: If $K \geq 2$, we start with $k = 1$. If j_k^* and j_{k+1}^* satisfy $j_{k+1}^* - j_k^* \leq 20$ and the signs of $\Delta X_{j_k^*}^*$ and $\Delta X_{j_{k+1}^*}^*$ are opposite, then we set all the asset prices from $t_{j_k^*}^*$ to $t_{j_{k+1}^*-1}^*$ equal to $S(t_{j_k^*}^*)$. After that, increase k by 2 and repeat Step 3. Otherwise, increase k by 1 and repeat Step 3. We stop the procedure as soon as the value of k becomes equal to K .

3 Detecting Common and Disjoint Jumps

We collected, cleaned, and sampled intraday data with a time interval of 1 sec for 251 trading days in 2018. The Ait-Sahalia and Jacod’s jump test [1] was implemented to detect jumps for each day. We considered two null hypotheses: (1) there is no jump; (2) there is at least one jump. If an intraday time series rejects the first null hypothesis and does not reject the second null hypothesis, the time series is considered to have jumps. For those dates when both the market index and an individual asset had jumps on the same day, the Jacod and Todorov’s common jump test [7] is used to verify the existence of common and disjoint jumps. Again, we consider two null hypotheses: (1) common jumps exist; (2) disjoint jumps exist. As a result, at the 95% confidence level, we have detected 14 days when both S&P500 and AAPL have jumps on the same day: 11 days with both common and disjoint types of jumps, 2 days with only common jumps, and 1 day with only disjoint jumps. Additionally, we found 18 days when both S&P500 and WMT have jumps on the same day: 11 days with both disjoint and common jumps, 6 days with disjoint jumps only, and 1 day with common jumps only. The corresponding p -values and acceptance results are presented in Table 1. This statistical analysis provides empirical evidence of the presence of disjoint and common jumps, implying that using jumps in a systemic component only is not sufficient to explain all jumps in the market dynamics for high-frequency data. Thus, it is reasonable to consider a multi-asset price model with common and asset-specific jumps.

4 Multi-Asset Price Model

Let S_0 denote a market index, which is considered as a systemic security affecting the dynamics of all other assets. Assume the market index value follows a jump-diffusion process with the following stochastic differential equation (SDE):

$$\frac{dS_0(t)}{S_0(t-)} = \mu_0 dt + \sigma_0 dW_0(t) + d\left(\sum_{\ell=1}^{N_{\lambda_0}(t)} (e^{\sigma_0 Q_\ell} - 1)\right), \quad t \geq 0. \tag{1}$$

Here, the constants μ_0 and $\sigma_0 > 0$ represent the drift rate and volatility of return of the market index, respectively; $\{W_0(t)\}_{t \geq 0}$ is a standard Brownian Motion; $\{Q_\ell\}$ is a sequence of iid normal random jump sizes with mean μ_{J_0} and variance $\sigma_{J_0}^2$; and $\{N_{\lambda_0}(t)\}_{t \geq 0}$ is a standard Poisson process with intensity $\lambda_0 > 0$. The usual assumption is that $\{Q_\ell\}$, $\{N_{\lambda_0}(t)\}$ and $\{W_0(t)\}$ are jointly independent. Obviously, if there are no jumps (e.g., $\lambda = 0$ or $\sigma_{J_0} = 0$), the solution to (1) is a Geometric Brownian Motion (GBM). We refer to it as the GBM case.

For the asset price model with common jumps only, we assume that the strong solution and the SDE are the same as those proposed in [4]:

Table 1 Jump test (JT) results for the index and one stock: p -values and acceptances

SPY and WMT			SPY and AAPL		
Date	Disjoint JT	Common JT	Date	Disjoint JT	Common JT
2018.02.26	0.849 ✓	0.000 ×	2018.01.23	0.562 ✓	0.146 ✓
2018.03.15	0.781 ✓	0.102 ✓	2018.02.15	0.565 ✓	0.083 ✓
2018.04.20	0.659 ✓	0.219 ✓	2018.03.15	0.172 ✓	0.267 ✓
2018.05.01	0.638 ✓	0.128 ✓	2018.05.01	0.139 ✓	0.406 ✓
2018.05.24	0.634 ✓	0.121 ✓	2018.05.25	0.010 ×	0.684 ✓
2018.06.12	0.725 ✓	0.371 ✓	2018.06.12	0.010 ×	0.727 ✓
2018.06.27	0.708 ✓	0.001 ×	2018.07.26	0.330 ✓	0.916 ✓
2018.07.26	0.404 ✓	0.797 ✓	2018.08.22	0.250 ✓	0.481 ✓
2018.08.28	0.768 ✓	0.000 ×	2018.08.23	0.210 ✓	0.560 ✓
2018.08.24	0.709 ✓	0.263 ✓	2018.08.29	0.471 ✓	0.689 ✓
2018.08.31	0.700 ✓	0.140 ✓	2018.09.21	0.121 ✓	0.799 ✓
2018.09.04	0.539 ✓	0.094 ✓	2018.09.14	0.369 ✓	0.390 ✓
2018.09.21	0.647 ✓	0.006 ×	2018.09.25	0.424 ✓	0.208 ✓
2018.10.02	0.838 ✓	0.002 ×	2018.11.16	0.165 ✓	0.030 ×
2018.11.01	0.630 ✓	0.650 ✓			
2018.11.30	0.028 ×	0.827 ✓			
2018.12.03	0.820 ✓	0.000 ×			
2018.12.21	0.457 ✓	0.066 ✓			

$$\frac{dS_i(t)}{S_i(t-)} = \mu_i dt + \sigma_i \left(\rho_i dW_0(t) + \sqrt{1 - \rho_i^2} dW_i(t) \right) + d \left(\sum_{\ell=1}^{N_{\lambda_0}(t)} (e^{\sigma_i \rho_i Q_\ell} - 1) \right), \tag{2}$$

for $t \geq 0$. Here μ_i and $\sigma_i > 0$ are, respectively, the drift rate and volatility for asset S_i , $i \geq 1$. The coefficient $\rho_i \in (-1, 1)$ represents the correlation between the systemic process S_0 and asset S_i . The Brownian motion $\{W_i(t)\}_{t \geq 0}$ is independent of $\{S_0(t)\}_{t \geq 0}$. Under the assumption that the market index has no jumps, the above SDE reduces to the GBM case.

Next, we propose a new multivariate jump-diffusion model, where the individual assets have not only jumps from the market index but also additional, asset-specific jumps independent from those of S_0 . The respective SDE is as follows:

$$\begin{aligned} \frac{dS_i(t)}{S_i(t-)} = & \mu_i dt + \sigma_i \left(\rho_i dW_0(t) + \sqrt{1 - \rho_i^2} dW_i(t) \right) \\ & + d \left(\sum_{\ell=1}^{N_{\lambda_0}(t)} (e^{\sigma_i \rho_i Q_\ell} - 1) + \sum_{k=1}^{N_{\lambda_i}(t)} (e^{\sigma_i \sqrt{1 - \rho_i^2} Q_k^{(i)}} - 1) \right), \quad t \geq 0. \end{aligned} \tag{3}$$

For each asset S_i with $i \geq 1$, the jump sizes $\{Q_k^{(i)}\}$ are iid random variables with mean μ_{J_i} and variance $\sigma_{J_i}^2$. Here, $\{N_{\lambda_i}(t)\}_{t \geq 0}$ is a Poisson process for jumps that are specific for asset S_i . All of $\{Q_k^{(i)}\}$, $\{N_{\lambda_i}(t)\}$ and $\{W_i(t)\}$ are independent from each other, as well as from the process $\{S_0(t)\}$. Additionally, conditional on $\{S_0(t)\}$, the price processes $\{S_i(t)\}$, $i \geq 1$ are jointly independent. Here, all jump sizes are normally distributed.

5 Estimation of Parameters for the Systemic Component

As described in [6], the multinomial maximum likelihood estimation (MMLE) method can be employed to speed up the calibration of the market index model. The MMLE method consists of the following two steps.

Step 1: Sort m historical log-returns into n_{bin} bins and find the *sample* frequency $f_k^{(s)}$ for each bin $k = 1, 2, \dots, n_{bin}$.

Step 2: Maximize the objective function: $l(\mathbf{v}) = \sum_{k=1}^{n_{bin}} [f_k^{(s)} \ln(f_k(\mathbf{v}))] \rightarrow \max_{\mathbf{v}}$.

The theoretical frequency $f_k(\mathbf{v})$ for the model distribution with probability density ϕ , cumulative distribution function Φ , and parameter vector \mathbf{v} is given by

$$f_k(\mathbf{v}) = m \int_{B_k} \phi(x; \mathbf{v}) dx = m(\Phi(b_k; \mathbf{v}) - \Phi(b_{k-1}; \mathbf{v})),$$

where $B_k = [b_{k-1}, b_k]$ is the k th bin.

Similarly to [4], a first-order approximation is used to calculate the theoretical frequencies. The probability of having k jumps during a time interval of length h is $p_k = e^{-\lambda h} (\lambda h)^k / k!$. The distribution of the log-return $\tilde{Z} = \ln(S_0(t+h)/S(t))$ from t to $t+h$ given that k jumps occur is normal with mean $(\mu_0 - \frac{\sigma_0^2}{2})h + k\mu_{J_0}$ and variance $\sigma_0^2 h + k\sigma_{J_0}^2$. Let $N_k(x)$ denote the cumulative distribution function (CDF) for this distribution. For small h , it is unlikely to have more than one jump per time period, and hence the CDF Φ of the log-return \tilde{Z} can be approximated as:

$$\Phi(x) \approx \frac{p_0}{p_0 + p_1} N_0(x) + \frac{p_1}{p_0 + p_1} N_1(x).$$

6 Estimation of Model Parameters

Our next step is to estimate the model parameters for each asset S_i with $i \geq 1$ by conditioning on the market index values. As demonstrated in [4], the model parameters can be individually estimated for each asset. The maximum likelihood method (MLE) is used for each of the two jump-diffusion asset price models described in Sect. 4.

Introduce the log-values $Z(t) = \ln S_0(t)$ and $X_i(t) = \ln S_i(t)$ for $i = 1, 2, \dots, n$ and $t \geq 0$. Fix the index $i \geq 1$ and use the following notations:

$$Z_j = \ln S_0(t_j), \quad X_j = \ln S_i(t_j), \quad M_j = W_0(t_j) + \sum_{k=1}^{N_{\lambda_0}(t_j)} Q_k,$$

where $j = 0, 1, \dots, m$, and m is the number of observations available. Additionally, we use lowercase letters z_j and x_j for historical values of log-prices.

6.1 Assets Model with only Common Jump

Consider the asset price model with only co-jumps. Find and reorganize the strong solutions to (1) and (2) to obtain:

$$X_j = X_{j-1} + \left(\mu_i - \frac{\sigma_i}{2} \right) h_j + \sigma_i \rho_i \left(\frac{\tilde{Z}_j - (\mu_0 - \frac{\sigma_0^2}{2}) h_j}{\sigma_0} \right) + \sigma_i \sqrt{1 - \rho_i^2} (W_i(t_j) - W_i(t_{j-1})),$$

where $h_j = t_j - t_{j-1}$, and $\tilde{Z}_j = Z_j - Z_{j-1}$ is the log-return $\ln(S_0(t_j)/S_0(t_{j-1}))$ for $j = 0, 1, 2, \dots, m$. The joint transitional PDF of X_j and Z_j is:

$$p_{X_j, Z_j | X_{j-1}, Z_{j-1}}(x_j, z_j | x_{j-1}, z_{j-1}) = p_{Z_j | Z_{j-1}}(z_j | z_{j-1}) \times p_{X_j | X_{j-1}, \tilde{Z}_j}(x_j | x_{j-1}, \tilde{z}_j),$$

where $\tilde{z}_j = z_j - z_{j-1}$.

We can construct a likelihood function conditional on values of S_0 (or its log-values Z). The transitional PDF $p_{Z_j | Z_{j-1}}(z_j | z_{j-1})$ can be omitted in the likelihood function. As a result, we have the following conditional log-likelihood function for asset S_i :

$$L_i(X | Z) = \prod_{j=1}^m p_{X_j | X_{j-1}, \tilde{Z}_j}(x_j | x_{j-1}, \tilde{z}_j). \tag{4}$$

Clearly, it is a product of normal densities.

To maximize the log-likelihood function in (4), we first find zeros of the partial derivatives w.r.t. the parameters μ_i , σ_i , and ρ_i . From [4], we have the following solution: $\sigma_i = \sqrt{u^2 + v^2}$, $\mu_i = w + \frac{u^2 + v^2}{2}$, and $\rho_i = \frac{v}{\sqrt{u^2 + v^2}}$ where

$$u^2 = \frac{\sum_{j=1}^m \left(\tilde{x}_j - h_j w - \frac{\tilde{z}_j}{\sigma_0} v + d_0 v h_j \right)^2 / h_j}{m},$$

$$v = \frac{(c - (\Delta x \Delta z / \Delta t)) \sigma_0}{b^2 - (\Delta z)^2 / \Delta t}, \quad w = \frac{\Delta x b^2 - \Delta z c + \tilde{\mu}_0 \Delta t c - \tilde{\mu}_0 \Delta x \Delta z}{\Delta t (b^2 - \Delta z / \Delta t)},$$

$d_0 = \frac{\mu_0 - \frac{\sigma_0^2}{2}}{\sigma_0}$, $\tilde{x}_j = x_j - x_{j-1}$, $\Delta t = \sum_{j=1}^m h_j = t_m - t_0$, $\Delta x = \sum_{j=1}^m \tilde{x}_j = x_m - x_0$, $\Delta z = \sum_{j=1}^m \tilde{z}_j = z_m - z_0$, and $\tilde{\mu}_0 = \mu_0 - \frac{\sigma_0^2}{2}$. The parameters b^2 and c are as follows:

$$b^2 = \sum_{j=1}^m \frac{\tilde{z}_j^2}{h_j} \text{ and } c = \sum_{j=1}^m \frac{\tilde{z}_j \tilde{x}_j}{h_j}.$$

6.2 Assets Model with an Additional Jump Component

Now, we consider the case with asset-specific jumps. Find and reorganize the strong solutions to (1) and (3) to obtain:

$$X_j = X_{j-1} + \left(\mu_i - \frac{\sigma_i}{2}\right)h_j + \sigma_i \rho_i \left(\frac{\tilde{Z}_j - (\mu_0 - \frac{\sigma_0^2}{2})h_j}{\sigma_0}\right) + \sigma_i \sqrt{1 - \rho_i^2} \left(W_i(t_j) - W_i(t_{j-1}) + \sum_{k=N_{\lambda_i}(t_{j-1})+1}^{N_{\lambda_i}(t_j)} Q_k^{(i)}\right). \tag{5}$$

There is an additional jump part in the equation, and the jump sizes for assets are normally-distributed. Assume there is at most one jump in each time interval $[t_{j-1}, t_j]$. In this case, the conditional distribution of X_j given X_{j-1} and $\tilde{Z}_j = Z_j - Z_{j-1}$ is a mixture of two normal distributions. The joint probability function of X_j and Z_j conditional on X_{j-1} and Z_{j-1} is as follows:

$$p_{X_j, Z_j | X_{j-1}, Z_{j-1}}(x_j, z_j | x_{j-1}, z_{j-1}) = p_{Z_j | Z_{j-1}}(z_j | z_{j-1}) \times p_{X_j | X_{j-1}, \tilde{Z}_j}(x_j | x_{j-1}, \tilde{z}_j),$$

$$p_{X_j | X_{j-1}, \tilde{Z}_j}(x_j | x_{j-1}, \tilde{z}_j) \approx \frac{p_0}{p_0 + p_1} \frac{1}{\sqrt{2\pi \sigma_i^2 h_j (1 - \rho_i^2)}} \times \exp \left\{ -\frac{\left[x_j - x_{j-1} - (\mu_i - \frac{\sigma_i^2}{2})h_j - \frac{\sigma_i \rho_i}{\sigma_0} \left(\tilde{z}_j - (\mu_0 - \frac{\sigma_0^2}{2})h_j \right) \right]^2}{2\sigma_i^2 (1 - \rho_i^2) h_j} \right\}$$

$$+ \frac{p_1}{p_0 + p_1} \frac{1}{\sqrt{2\pi \sigma_i^2 (h_j + \sigma_{J_i}^2) (1 - \rho_i^2)}} \times \exp \left\{ -\frac{\left[x_j - x_{j-1} - (\mu_i - \frac{\sigma_i^2}{2})h_j - \frac{\sigma_i \rho_i}{\sigma_0} \left(\tilde{z}_j - (\mu_0 - \frac{\sigma_0^2}{2})h_j \right) - \sigma_i \sqrt{1 - \rho_i^2} \mu_{J_i} \right]^2}{2\sigma_i^2 (1 - \rho_i^2) (h_j + \sigma_{J_i}^2)} \right\}.$$

Again, when we construct the conditional log-likelihood function $L_i(X | Z)$, we omit the density function $p_{Z_j | Z_{j-1}}(z_j | z_{j-1})$. It is too complicated to find zeros of derivatives of the likelihood function in closed-form due to the inclusion of additional

jumps. Instead, we can use a numerical optimization method such as the Nelder–Mead simplex method available in MATLAB under the function *fminsearch*.

We assume that the systemic component models the market index and that other assets depend on it. The same idea is used in the Capital Asset Price Model (CAPM). This approach simplifies the calibration procedure: estimate the parameters of the systemic component first and then find the parameters of every other asset price process. The other possible approach is to treat the systemic asset as an unobservable, hidden process like in the Hidden Markov Model (HMM). The market index can then be regarded as one of the assets available on the market. However, the calibration of such a multi-asset model is not a straightforward task.

7 Numerical Results

In this paper, we calibrated all models using: (1) the intraday data with the time interval equal to 1 s and (2) the intraweek data with the time interval equal to 5 minutes. The parameters for the GBM market index model (without jumps) are used as the initial guess for the MMLE method when we estimate parameters of the jump-diffusion market index model. For individual assets, we first estimate the parameters for the model without asset-specific jumps using analytical formulae from Sect. 6.1. The estimation process is conditional on the market index values. Secondly, to estimate parameters for the asset price model with asset-specific jumps, we use the results for the model with common jumps only as an initial approximation. Table 2 shows calibration results for the intraday data on 15 March 2018 with time intervals equal to 1 s. The jump detection tests indicate that the last week in August 2018 appears to have more jumps. We re-sampled intraday data from 27 August to 31 August 2018 with a 5-minute time step and merged daily data with log-returns. The intraweek parameter estimation results are presented in Table 3.

We employed the Kullback–Leibler (KL) divergence to measure how close the probability distribution for each model to the historical distribution. We measure the empirical CDF, denoted by F , using the function *ecdf* in MATLAB. Meanwhile, Q denotes the CDF for one of the asset price models considered in this paper. For discrete probability distributions F and Q , the KL divergence from F to Q denoted as $D_{KL}(Q\|F)$ and from Q to F denoted as $D_{KL}(F\|Q)$. As the KL divergence measure is not symmetric, we compute both $D_{KL}(Q\|F)$ and $D_{KL}(F\|Q)$, then sum them together, where

$$D_{KL}(Q\|F) = \sum_{x \in X} Q(x) \ln \left(\frac{Q(x)}{F(x)} \right) \text{ and } D_{KL}(F\|Q) = \sum_{x \in X} F(x) \ln \left(\frac{F(x)}{Q(x)} \right).$$

The Kolmogorov–Smirnov (KS) test is used to measure the goodness of fit of the models. The null hypothesis of the KS test is that the sample is drawn from the reference distribution. The KS statistic quantifies a distance between the empirical

Table 2 Daily parameters for $\Delta t = 1$ sec

Market index		μ_0	σ_0		μ_{J_0}	σ_{J_0}	λ_0
SPY	GBM	-0.0029	0.0067				
	Jump-diffusion	-0.0022	0.0061		0.0025	0.0451	1.0165
Assets		μ_i	σ_i	ρ_i	μ_{J_i}	σ_{J_i}	λ_i
WMT	GBM	-0.0004	0.0187	0.1372			
$i = 1$	With co-jumps only	-0.0001	0.0187	0.1251			
	With disjoint jumps	0.00001	0.0129	0.1516	-0.00002	0.0007	1.0027
AAPL	GBM	0.0007	0.0101	0.3801			
$i = 2$	With co-jumps only	0.0011	0.0099	0.3492			
	With disjoint jumps	0.0006	0.0093	0.3593	-0.000004	0.0004	1.0082

Table 3 Weekly parameters for $\Delta t = 5$ min

Market Index		μ_0	σ_0		μ_{J_0}	σ_{J_0}	λ_0
SPY	GBM	0.0051	0.0078				
	Jump-diffusion	0.0078	0.0070		-0.1525	0.00004	0.9984
Assets		μ_i	σ_i	ρ_i	μ_{J_i}	σ_{J_i}	λ_i
WMT	GBM	0.0026	0.0174	0.3357			
$i = 1$	With co-jumps	0.0047	0.0172	0.3053			
	With disjoint jumps	-0.0037	0.0161	0.3192	0.0040	0.0012	0.9916
AAPL	GBM	0.0304	0.0230	0.5426			
$i = 2$	With co-jumps	0.0347	0.0223	0.5024			
	With disjoint jumps	0.0396	0.0202	0.533	-0.0010	0.00383	1.004

Table 4 KL divergence & KS statistic values for intraday data with $\Delta t = 1$ sec

	Market index	SPY	Assets	WMT	AAPL
KL	GBM	157.4857	GBM	480	220.4032
Divergence	Jump-diffusion	131.0703	With co-jumps	477.8103	209.8041
			With disjoint jumps	280.644	184.7301
KS	GBM	0.0544	GBM	0.047	0.0387
Statistics	Jump-diffusion	0.0696	With co-jumps	0.047	0.0419
			With disjoint jumps	0.1004	0.0516

Table 5 KL divergence & KS statistic values for intraweek data with $\Delta t = 5$ min

	Market index	SPY	Assets	WMT	AAPL
KL	GBM	1.6253	GBM	1.1225	2.1719
Divergence	Jump-diffusion	0.786	With co-jumps	1.0093	1.6165
			With disjoint jumps	0.5462	0.79
KS	GBM	0.0703	GBM	0.0596	0.0695
Statistics	Jump-diffusion	0.0449	With co-jumps	0.0595	0.0676
			With disjoint jumps	0.0424	0.0563

distribution function of the sample and the CDF of the reference distribution. Let the empirical distribution function be denoted by F . The KS statistic for a given reference CDF Q is $D_{KS} = \sup_x |F(x) - Q(x)|$.

The CDF of the market index model can be derived directly. The unconditional CDF for each asset price models can be obtained using properties of the normal distribution. As a result, we have derived the unconditional CDF of the i th asset’s log-return over a time interval of length h for the model with common and asset-specific jumps as a linear combination of normal CDFs.

We compared two distributions of the market index log-returns: the GBM and jump-diffusion models, and three distributions of the asset log-returns: the GBM, jump-diffusion with commons jumps only, and jump-diffusion with additional asset-specific jumps. The KL divergence values and KS test statistic results are reported in Tables 4 and 5.

Firstly, we calculated the KL divergence for the market index and assets. A smaller KL divergence indicates a better model that fits the data. The results reported in two tables support that the jump-diffusion model for the market index outperforms the

GBM process without jumps. Secondly, we compared models by KS statistics. A smaller KS statistic value indicates better goodness of fit of the model. The results reported in Table 4 show that the GBM model without jumps is the best fit for the observations for both market index and assets when time interval equals to 1 s. Table 5 shows the KS statistic values at the 5-minute time level. The presented results support that the jump-diffusion model for the market index fits the empirical data better than the GBM model without jumps. In summary, the jump-diffusion asset price model with both common and asset-specific types of jumps is the best fit among the three asset price models considered in this paper.

Acknowledgements R. Makarov wishes to acknowledge the support of the NSERC Discovery Grant program.

References

1. Aït-Sahalia, Yacine; Jacod, Jean: Testing for jumps in a discretely observed process. *Ann. Stat.* **37**(1), 184–222 (2009)
2. Aït-Sahalia, Y.; Jacod, J.: *High-Frequency Financial Econometrics*. Princeton University Press, Princeton (2014)
3. Aït-Sahalia, Y., Jacod, J., et al.: Is Brownian motion necessary to model high-frequency data? *Ann. Stat.* **38**(5), 3093–3128 (2010)
4. Chen, Y., Makarov, R.N.: Modelling asynchronous assets with jump-diffusion processes. In: *International Conference on Applied Mathematics, Modeling and Computational Science*, pp. 477–487. Springer (2017)
5. Cont, Rama: Statistical modeling of high-frequency financial data. *IEEE Sig. Process. Mag.* **28**(5), 16–25 (2011)
6. Hanson, F.B., Westman, J.J., Zhu, Z.: Multinomial maximum likelihood estimation of market parameters for stock jump-diffusion models. *Contemp. Math.* **351**, 155–170 (2004)
7. Jacod, Jean; Todorov, Viktor: Testing for common arrivals of jumps for discretely observed multidimensional processes. *Ann. Stat.* **37**(4), 1792–1838 (2009)
8. Jing, B.-Y., Kong, X.-B., Liu, Z., et al.: Modeling high-frequency financial data by pure jump processes. *Ann. Stat.* **40**(2), 759–784 (2012)
9. Lee, S.S., Hannig, J.: Detecting jumps from Lévy jump diffusion processes. *J. Financ. Econ.* **96**(2):271–290 (2010)