Mathematics in Industry 39
The European Consortium for Mathematics in Industry

Matthias Ehrhardt
Michael Günther *Editors*

Progress in Industrial Mathematics at ECMI 2021

ECMI
EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

Springer

MATHEMATICS IN INDUSTRY 39

THE EUROPEAN CONSORTIUM
FOR MATHEMATICS IN INDUSTRY

SUBSERIES

Mathematics in Industry focuses on the research and educational aspects of mathematics used in industry and other business enterprises. Books for *Mathematics in Industry* are in the following categories: research monographs, problem-oriented multi-author collections, textbooks with a problem-oriented approach, and conference proceedings. All manuscripts undergo rigorous peer review before acceptance. Relevance to the actual practical use of mathematics in the industry is the distinguishing feature of the books in the *Mathematics in Industry* series.

Matthias Ehrhardt • Michael Günther
Editors

Progress in Industrial Mathematics at ECMI 2021

21st ECMI Conference, Wuppertal, Germany, April 13–15, Selected and reviewed papers

Springer

ECMI
EUROPEAN CONSORTIUM FOR
MATHEMATICS IN INDUSTRY

Editors

Matthias Ehrhardt
Angewandte Mathematik und Numerische Analysis
Bergische Universität Wuppertal
Wuppertal, Germany

Michael Günther
Angewandte Mathematik und Numerische Analysis
Bergische Universität Wuppertal
Wuppertal, Germany

ISSN 1612-3956 ISSN 2198-3283 (electronic)
Mathematics in Industry
The European Consortium for Mathematics in Industry
ISBN 978-3-031-11817-3 ISBN 978-3-031-11818-0 (eBook)
https://doi.org/10.1007/978-3-031-11818-0

Mathematics Subject Classification: 00A69, 35-XX, 49-XX, 65-XX

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

Foreword

The proceedings of the ECMI 2021 conference, Progress in industrial mathematics 2022, contains contributions from many areas of mathematics relevant to social and economic development. The conference was hosted by the University of Wuppertal, Germany, and we can be proud of the success of the conference in difficult pandemic times. Although it was held online, all the important features of traditional ECMI conferences were maintained, starting with plenary lectures, award-winning talks, mini-symposia on interesting topics from ECMI Special Interest Groups, and many contributed papers. We were pleased to welcome colleagues from the Asia Pacific Consortium for Mathematics in Industry to our conference.

Considering the reality in 2020–2022 and the importance of mathematical modeling in times of pandemics, a number of papers in this conference volume deal with various epidemiological models for diseases caused by viruses, followed by industry-related models such as fluid dynamics, modeling of various industrial processes, problems from agriculture, modeling of various financial products, etc. A number of papers in this book deal with data analysis and data-driven models. Thus, the book provides a relevant overview of the topics and advances in industrial mathematics at the present time.

Educating industrial mathematicians is one of ECMI's main activities, and our signature events are the Mathematical Modeling Weeks. Last year, the Mathematical Modeling Week was held for the first time in Russia and for the first time online. A report on this event is included in this conference proceedings.

The list of authors includes a number of PhD students representing a new generation of industrial mathematicians trained in close collaboration with industrial partners under the Marie Skłodowska-Curie European Industrial PhD programs.

On behalf of ECMI and all conference participants, I would like to express our gratitude to Prof. Matthias Ehrhardt and his team for organizing the conference.

Novi Sad, Serbia
February 2022

Nataša Krejić, ECMI President

Preface

The 21st European Conference on Mathematics for Industry, ECMI 2021, was held online (organized by the University of Wuppertal) from 13 to 15 April 2021, bringing together more than 350 researchers for intellectual interaction for 3 days.

ECMI

EUROPEAN CONSORTIUM FOR
MATHEMATICS IN INDUSTRY

2021
CONFERENCE

The European Consortium for Mathematics in Industry (ECMI) organized its first international conference in Oberwolfach, in 1983, followed by a series of conferences, a persistent objective of which has been to galvanize interaction between academy and industry, leading to innovations in both fields. The 21st conference, ECMI 2021, inspired multidisciplinary research along these lines further, leading to the formulation of real-life challenges, where mathematical technologies provided significant new insights. Following the traditions of ECMI, the conference focused on various fields of industrial and applied mathematics, such as Applied Physics, Biology and Medicine, Cybersecurity, Data Science, Economy, Finance and Insurance, Energy, Production Systems, Social Challenges, Vehicles, and Transportation. These themes nicely fit to current distinguished national research programs in Hungary, in particular programs on Autonomous Vehicles, Digital Factories, Brain Research, or Precision Agriculture, supported by the EU and the National Research, Development and Innovation Office.

This virtual conference was organized by the University of Wuppertal and implemented using the ZOOM online conference system. The statistics of the conference were more than satisfactory. In addition to the 14 plenary talks, given by world class researchers and the three awardees, we had 28 minisymposia, and 24 contributed talks, running in up to 10 parallel sessions. Altogether there were more than 350 participants, from around 40 countries. More than 160 participants were students.

The Scientific Committee was set up as follows:

Adérito Araújo, University of Coimbra, Portugal
Linda Cummins, New Jersey Institute of Technology, USA
Michel Destrade, National University of Ireland Galway, Ireland
Matthias Ehrhardt, Bergische Universität Wuppertal, Germany (Chair)
Andrew Fowler, University of Oxford, UK
James Gleeson, University of Limerick, Ireland
Michael Günther, Bergische Universität Wuppertal, Germany (Co-Chair)
Helge Holden, NTNU Trondheim, Norway
Barbara Kaltenbacher, Alpen-Adria Universität, Klagenfurt, Austria
Nataša Krejić, University of Novi Sad, Serbia
Stig Larsson, Chalmers University of Technology, Sweden
Tim Myers, Universitat Autonoma de Barcelona, Spain
Stephen O'Brien, University of Limerick, Ireland
Jörg Osterrieder, ZHAW School of Engineering, Winterthur, Switzerland
René Pinnau, TU Kaiserslautern, Germany
Peregrina Quintela Estévez, University of Santiago de Compostela, Spain
Claudia Sagastizabal, UNICAMP and IMPA, Brasil
Sarah Waters, University of Oxford, UK

The local Organizing Committee in Wuppertal was

Markus Bannenberg, Bergische Universität Wuppertal
Andreas Bartel, Bergische Universität Wuppertal
Anna Clevenhaus, Bergische Universität Wuppertal
Matthias Ehrhardt, Bergische Universität Wuppertal (Chair)
Stephanie Friedhoff, Bergische Universität Wuppertal
Michael Günther, Bergische Universität Wuppertal
Jens Jäschke, Bergische Universität Wuppertal
Lorenc Kapllani, Bergische Universität Wuppertal
Jörg Kienitz, Bergische Universität Wuppertal
Friedemann Klaß, Bergische Universität Wuppertal
Tatiana Kossaczká, Bergische Universität Wuppertal
Michelle Muniz, Bergische Universität Wuppertal
Hanna Schilar, Bergische Universität Wuppertal (Conference Secretary)
Long Teng, Bergische Universität Wuppertal
Jan ter Maten, Eindhoven
Renate Winkler, Bergische Universität Wuppertal
Mario Zaghini, Bergische Universität Wuppertal

The plenary talks covered several major areas of applied and industrial mathematics, such as network theory, numerical methods of PDEs, mathematics of tomography, mechanical models, traffic management, control theory, cancer research, and environmental modeling. The plenary speakers were:

Nira Chamberlain, Loughborough University
Andrea Bertozzi, University of California Los Angeles
Elisabetta Rocca, University of Pavia
Carola-Bibiane Schönlieb, University of Cambridge
Eddie Wilson, University of Bristol
Mark McGuinness, Victoria University of Wellington
Luis Nunes Vicente, Lehigh University and CMUC-Portugal
Ginestra Bianconi, Queen Mary University of London
Anna-Karin Tornberg, KTH Stockholm
William Lee, University of Huddersfield
Paul Dellar, University of Oxford

The winner of the Anile prize, honoring Professor Angelo Marcello Anile (1948–2007) of the University of Catania, is given to a young researcher for an excellent PhD thesis in industrial mathematics. The Anile prize, in 2021, was awarded to Bernadette Stolz (University of Oxford) ("Global and local persistent homology for the shape and classification of biological data").

The Hansjörg Wacker Memorial Prize, established in memory of ECMI founding member Hansjörg Wacker, (1939–1991), who was Professor at the Johannes Kepler University, Linz, is awarded for the best mathematical dissertation at the Master's level on an industrial project. The Hansjörg Wacker Memorial Prize, in 2021, was awarded to Halvor Snersrud Gustad (NTNU Trondheim) ("Using Artificial Neural Networks for Predicting Bending Moments of Riser Structures") and Jan Brekelmans (TU Eindhoven) ("The Volume-of-Fluid Method Applied to Vertical Slug Flow Using an Axial-Symmetric and a Fully Three-Dimensional Approach").

The Organizers express their deepest gratitude to everybody involved in the success of this meeting, the plenary speakers, the members of the Scientific Committee, the organizers of the minisymposia, the contributing authors, and all the participants of the conference.

On behalf of the Organizers

Wuppertal, Germany	Matthias Ehrhardt
Wuppertal, Germany	Michael Günther
February 2022	

Contents

Contributors

Parveena Shamim Abdul Salam Technische Universität Kaiserslautern, Kaiserslautern, Germany

Maria Aguareles Universitat de Girona, Girona, Spain

Fawaz K. Alalhareth The University of Texas at Arlington, Department of Mathematics, Arlington, TX, USA
Najran University, Department of Mathematics, Najran, Saudi Arabia

Miracle Amadi LUT School of Engineering Science, LUT University, Lappeenranta, Finland

Adérito Araújo CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal

Patrícia Araújo Nors Group, Porto, Portugal

Martin Arnold Institute of Mathematics, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Iñigo Arregui CITIC and University of A Coruña, A Coruña, Spain

Milton Assunção Mathematics Applications Consortium for Science and Industry (MACSI), Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

A. M. C. H. Attanayake Department of Statistics and Computer Science, University of Kelaniya, Kelaniya, Sri Lanka

Elisa Atza Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

Martine J. Barons AS&RU, Department of Statistics, University of Warwick, Coventry, UK

Sofia Barroso Nors Group, Porto, Portugal

Flavia Barsotti ING Analytics, Amsterdam, The Netherlands
IAS (Institute for Advanced Study), University of Amsterdam, Amsterdam, The Netherlands

Andreas Bartel University of Wuppertal, Wuppertal, Germany

Benjamin Bauer Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany

Klaus-Dieter Bauer MathConsult GmbH, and Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

Peter Benner Computational Methods in Systems and Control Theory Group at the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Jacques Besson Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Andreas Binder MathConsult GmbH, Linz, Austria

Samy Blusseau Mines Paris, Centre de Morphologie Mathématique, Fontainebleau, France

Wolfgang Bock TU Kaiserslautern, Kaiserslautern, Germany

Lívia Boda Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary

Laury-Hann Brassart Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Neil Budko Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

Jan Pablo Burgard Universität Trier, Trier, Germany

Lilli Burger Division Mathematics for Vehicle Engineering, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany

Michael Burger Division Mathematics for Vehicle Engineering, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany

Marc Calvo-Schwarzwalder Zayed University, Abu Dhabi, UAE

Vito Dario Camiola University of Catania, Catania, Italy

Tiago Carmo Nors Group, Porto, Portugal

Ana Carpio Universidad Complutense de Madrid, Madrid, Spain

Sabrina Casper Fresenius Medical Care Deutschland GmbH, Bad Homburg, Germany

Elena Celledoni Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Kurt Chudej Universität Bayreuth, Lehrstuhl für Wissenschaftliches Rechnen, Bayreuth, Germany

Anna Clevenhaus Bergische Universität Wuppertal, Wuppertal, Germany

Alberto Coccarelli Swansea University, Swansea, UK

Ergys Çokaj Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Juancho A. Collera Department of Mathematics and Computer Science, University of the Philippines Baguio, Baguio City, Philippines

Graziana Colonna CITIC and University of A Coruña, A Coruña, Spain

Manuel Cruz LEMA - Engineering Mathematics Laboratory, School of Engineering, Polytechnic of Porto, Porto, Portugal

Rainier Ric B. de la Cruz Department of Economics and Political Science, University of the Philippines Baguio, Baguio City, Philippines

Paul J. Dellar Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Oxford, UK

Francesco Delloro Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Vanessa Dörlich Fraunhofer ITWM, Kaiserslautern, Germany

Bertram Düring Mathematics Institute, University of Warwick, Coventry, UK

Pravir K. Dutt Department of Mathematics, IIT Kanpur, Kanpur, India

Fredrik Edelvik Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden

Matthias Ehrhardt Angewandte Mathematik und Numerische Analysis, Bergische Universität Wuppertal, Wuppertal, Germany

Hasitha Erandi Department of Mathematics, University of Colombo, Colombo, Sri Lanka

Lars Erhardsson Scania CV AB, Södertälje, Sweden

István Faragó Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary

Ana M. Ferreiro-Ferreiro CITIC and University of A Coruña, A Coruña, Spain

Gaby Folger Universität Bayreuth, Lehrstuhl für Wissenschaftliches Rechnen, Bayreuth, Germany

Thais C. O. Fonseca Department of Statistical Methods, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Francesc Font Universitat Politècnica de Catalunya, Barcelona, Spain

Mads Fromreide NORCE Norwegian Research Centre, Kristiansand, Norway

Doris H. Fuertinger Fresenius Medical Care Deutschland GmbH, Bad Homburg, Germany

Mona Fuhrländer Computational Electromagnetics Group (CEM) and Centre for Computational Engineering (CCE), TU Darmstadt, Darmstadt, Germany

Alessandro Gabbana Eindhoven University of Technology, Eindhoven, The Netherlands

Naleen Ganegoda Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka

Perfect Y. Gidisu TU Eindhoven, Eindhoven, The Netherlands

Michele Girfoglio Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

Rafael González-Albaladejo Universidad Complutense de Madrid, Madrid, Spain
Instituto Gregorio Millán, Universidad Carlos III de Madrid, Madrid, Spain

Simone Göttlich University of Mannheim, School of Business Informatics and Mathematics, Mannheim, Germany

Thomas Götz Mathematical Institute, University of Koblenz-Landau, Campus Koblenz, Koblenz, Germany

Anne-Françoise Gourgues-Lorenzon Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Sara Grundel Computational Methods in Systems and Control Theory Group at the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Olivier Guéant Université Paris 1 Panthéon Sorbonne, Centre d’Economie de la Sorbonne, Paris, France

Michael Günther Angewandte Mathematik und Numerische Analysis, Bergische Universität Wuppertal, Wuppertal, Germany

Heikki Haario LUT School of Engineering Science, LUT University, Lappeenranta, Finland

Svenn Anton Halvorsen NORCE Norwegian Research Centre, Kristiansand, Norway

Monika Harant Mathematics for the Digital Factory, Fraunhofer ITWM, Kaiserslautern, Germany

Marc Harmening Trier University, Trier, Germany

Josef Haslinger MathConsult GmbH, and Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

Sarah-Alexa Hauschild Universität Trier, Trier, Germany

Peter Heidrich Mathematical Institute, University of Koblenz-Landau, Campus Koblenz, Koblenz, Germany
Magister Laukhard IGS Herrstein-Rhaunen, Herrstein, Germany

Matti Heiliö Computational and Process Engineering, LUT University, Lappeenranta, Finland

Christof Heuer d-fine GmbH, Frankfurt, Germany

Stefan Heyder Technische Universität Ilmenau, Ilmenau, Germany

Christian Himpe Computational Methods in Systems and Control Theory Group, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Michiel E. Hochstenbach TU Eindhoven, Eindhoven, The Netherlands

Peter Holliman Swansea University, Swansea, UK

Thomas Hotz Technische Universität Ilmenau, Ilmenau, Germany

Birgit Jacob Bergische Universität Wuppertal, Wuppertal, Germany

Onkar Jadhav Institut für Mathematik, Berlin, Germany

Hamid Tamaddon Jahromi Swansea University, Swansea, UK

Jens Jäschke Bergische Universität Wuppertal, Wuppertal, Germany

Michel Jeandin Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Tomas Johnson Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden

Jason Jones Swansea University, Swansea, UK

Hiroyuki Kaimori Science Solutions International Laboratory, Inc., Tokyo, Japan

Akihisa Kameari Science Solutions International Laboratory, Inc., Tokyo, Japan

Lorenc Kapllani Bergische Universität Wuppertal, Wuppertal, Germany

Chris Kershaw Swansea University, Swansea, UK

Axel Klar Technische Universität Kaiserslautern, Kaiserslautern, Germany

Friedemann Klass University of Wuppertal, Wuppertal, Germany

Hristo V. Kojouharov The University of Texas at Arlington, Department of Mathematics, Arlington, TX, USA

Michael Kolmbauer MathConsult GmbH, Linz, Austria

Tatiana Kossaczká Angewandte Mathematik und Numerische Analysis, Bergische Universität Wuppertal, Wuppertal, Germany

Peter Kotanko Renal Research Institute New York, New York, NY, USA

Tsvetan Kotsev Sofia University, Department of Mathematics and Informatics, Sofia, Bulgaria

Tyll Krueger Wroclaw University of Science and Technology, Wrocław, Poland

V. K. Kukreja Department of Mathematics, SLIET Longowal, Longowal, Punjab, India

Archna Kumari Department of Mathematics, SLIET Longowal, Longowal, Punjab, India

Andrea Leone Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Joachim Linn Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany

U. P. Liyanage Department of Statistics and Computer Science, University of Kelaniya, Kelaniya, Sri Lanka

Sergey Lupuleac Institute of Applied Mathematics and Mechanics, Peter the Great St.Petersburg Polytechnic University, St Petersburg, Russia

Mahdi Hedayat Mahmoudi Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Benjamin Maldon School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia

Davide Manfredo Fraunhofer ITWM, Kaiserslautern, Germany

Nicole Marheineke Universität Trier, Trier, Germany

Andreas Mark Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden

Luca Mechelli Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany

Volker Mehrmann Institut für Mathematik, Berlin, Germany

Hannah Merwood Government Operational Research Service, London, UK

Takeshi Mifune Graduate School of Engineering, Kyoto University, Kyoto, Japan

Miklós E. Mincsovics Department of Differential Equations at Budapest University of Technology and Economics, Budapest, Hungary
Large Networks Research Group, ELKH, Budapest, Hungary

Katja Mombaur Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON, Canada

Rahele Mosleh Budapest University of Technology and Economics, Budapest, Hungary

Michelle Muniz Bergische Universität Wuppertal, Wuppertal, Germany

Davide Murari Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Timothy G. Myers Centre de Recerca Matemàtica, Barcelona, Spain

Matthias B. Näf Department of Mechanical Engineering, Vrije Universiteit Brussel, Brussels, Belgium

Giovanni Nastasi University of Catania, Catania, Italy

Perumal Nithiarasu Swansea University, Swansea, UK

Günter Offner AVL List GmbH, Graz, Austria

Brynjulf Owren Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Tigran Parikyan AVL List GmbH, Graz, Austria

S. S. N. Perera Research and Development Centre for Mathematical Modelling, Faculty of Science, University of Colombo, Colombo, Sri Lanka

Bernhard Pöchtrager Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria

Tatiana Pogarskaia Institute of Applied Mathematics and Mechanics, Peter the Great St.Petersburg Polytechnic University, St Petersburg, Russia

Roland Pulch Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany

Sandra Ramos LEMA - Engineering Mathematics Laboratory, School of Engineering, Polytechnic of Porto, Porto, Portugal

Vetle Kjær Risinggård NORCE Norwegian Research Centre, Kristiansand, Norway

Jan Rohleff Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany

Gilles Rolland EDF-Lab Les Renardières, Matériaux et Mécanique des Composants, Moret-sur-Loing Cedex, France

Sam Rolland Swansea University, Swansea, UK

Michael Roller Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany

Vittorio Romano University of Catania, Catania, Italy

Gianluigi Rozza Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

Sunčica Sakić Department of Numerical Mathematics, Charles University, Prague, Czech Republic

Niklas Sandgren IPS IBOFlow AB, Gothenburg, Sweden

Simon Sandgren Scania CV AB, Södertälje, Sweden

Jorge Santos LEMA - Engineering Mathematics Laboratory, School of Engineering, Polytechnic of Porto, Porto, Portugal

Igor Sazonov Swansea University, Swansea, UK

Fabio Schneider Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany

Falco Schneider Fraunhofer ITWM, Kaiserslautern, Germany

Sebastian Schöps Computational Electromagnetics Group (CEM) and Centre for Computational Engineering (CCE), TU Darmstadt, Darmstadt, Germany

Gabriella Svantnerné Sebestyén Budapest University of Technology and Economics, Budapest, Hungary

Nirav Vasant Shah Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

Shallu Shallu Department of Mathematics, SLIET Longowal, Longowal, Punjab, India

Arsha Sherly TU Kaiserslautern, Kaiserslautern, Germany

Bernd Simeon Felix-Klein Zentrum, TU Kaiserslautern, Kaiserslautern, Germany

Roberta Simonella CITIC and University of A Coruña, A Coruña, Spain

Peeyush Singh Department of Mathematics, VIT-AP University, Amaravati, Andhra Pradesh, India

Angela Slavova Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Manuel Sparta NORCE Norwegian Research Centre, Kristiansand, Norway

Nadine Stahl Trier University, Trier, Germany

Ngamta Thamwattana School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia

Hywel Thomas Swansea University, Swansea, UK

Sudarshan Tiwari Technische Universität Kaiserslautern, Kaiserslautern, Germany

Chedly Tizaoui Swansea University, Swansea, UK

Claudia Totzeck University of Wuppertal, School of Mathematics and Natural Sciences, Wuppertal, Germany

Zhomart Turarov TU Kaiserslautern, Kaiserslautern, Germany

David H. Underdown The National Archives, Surrey, UK

Abel Valverde Centre de Recerca Matemàtica, Barcelona, Spain

Carlos Vázquez CITIC and University of A Coruña, A Coruña, Spain
ITMATI, Campus Vida, Santiago de Compostela, Spain

Giorgia Vitanza University of Catania, Catania, Italy

Stefan Volkwein Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany

Michael Vynnycky Mathematics Applications Consortium for Science and Industry (MACSI), Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

W. A. U. K. Wetthasinghe Research & Development Centre for Mathematical Modelling, Department of Mathematics, University of Colombo, Colombo, Sri Lanka

Karunia Putra Wijaya Mathematical Institute, University of Koblenz, Koblenz, Germany

François Willot Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France
Mines Paris, Centre de Morphologie Mathématique, Fontainebleau, France

Jochen Wittmann Environmental Informatics, University of Applied Sciences (HTW) Berlin, Berlin, Germany

Raimund Wegener Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), Kaiserslautern, Germany

David Worsley Swansea University, Swansea, UK

Barbara Zubik-Kowal Department of Mathematics, Boise State University, Boise, ID, USA

Model Reduction for a Port-Hamiltonian Formulation of the Euler Equations

Sarah-Alexa Hauschild and Nicole Marheineke

Abstract The port-Hamiltonian (pH) formulation of partial-differential equations and their numerical treatment have been elaborately studied lately. This energy-based formulation encodes physical principles in the system structure and the pH-character is inherited during coupling. Considering a non-isothermal compressible fluid flow in a pipe, we propose a pH-model on PDE-level, which is advantageous for structure-preserving approximations. Based on Galerkin projection with compatible finite dimensional spaces we preserve the pH-structure during space discretization and model reduction. Numerical results support our theoretical findings.

1 Model Problem

This work focuses on the non-isothermal compressible Euler equations with friction and cooling on $\omega = [0, L]$ for $t \geq 0$,

$$\begin{aligned} 0 &= \partial_t \rho + \partial_x(\rho v), \\ 0 &= \partial_t(\rho v) + \partial_x(\rho v^2) + \partial_x p + \frac{\lambda}{2d}\rho|v|v, \\ 0 &= \partial_t e + \partial_x(ev) + p\partial_x v - \frac{\lambda}{2d}\rho|v|v^2 + \frac{4k_\omega}{d}(T - T_\infty). \end{aligned} \tag{1}$$

System (1) describes a fluid flow through a pipe with length L, diameter d, friction factor λ and heat transfer coefficient k_ω. The ambient temperature is denoted by T_∞. The unknowns are the mass density ρ, the velocity v and the internal energy density e. The system is closed with state equations for pressure p and temperature

S.-A. Hauschild (✉) · N. Marheineke
Universität Trier, Trier, Germany
e-mail: hauschild@uni-trier.de; marheineke@uni-trier.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_1

T as well as appropriate initial and boundary conditions. The Hamiltonian is given by $\mathcal{H}(\rho, v, e) = \int_0^L \rho\frac{v^2}{2} + e\,\mathrm{d}x$. In the following, $m = \rho v$ denotes the mass flow.

2 Port-Hamiltonian Formulation and Approximation

Following the ideas of [5] for finite dimensional pH-systems, we set up a non-unique pH-formulation for the infinite dimensional System (1).

A port-Hamiltonian formulation of (1) is given by

$$E(z)\partial_t z = (J(z) - R(z))\tilde{e}(z) + Gu(t) \tag{2}$$

with $z = (\rho, m, e)^T$ and $\tilde{e}(z) = (\frac{m^2}{2\rho^2}, m, 1)^T$ and system operators

$$E(z) = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{m}{\rho^2} & \frac{1}{\rho} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R(z) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{4k_\omega}{d}T \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ 0 \\ \frac{4k_\omega}{d} \end{pmatrix},$$

$$J(z) = \begin{pmatrix} 0 & -D_x & 0 \\ -D_x & 0 & -\frac{\lambda m|m|}{2d\rho^2} - \frac{e}{\rho}D_x - \frac{1}{\rho}D_x p \\ 0 & \frac{\lambda m|m|}{2d\rho^2} - D_x\frac{e}{\rho} - pD_x\frac{1}{\rho} & 0 \end{pmatrix}. \tag{3}$$

The input is given as $u(t) = T_\infty$. The Hamiltonian is $\mathcal{H}(z) = \int_0^L \frac{m^2}{2\rho} + e\,\mathrm{d}x$.

The operator D_x in (3) is defined as $(g_1 D_x g_2)g_3 := g_1\partial_x(g_2 g_3)$ for functions g_1, g_2 und g_3. System (2) fulfills the condition $E(z)^T\tilde{e}(z) = \frac{\delta\mathcal{H}}{\delta z}(z)$ and the operator $J(z)$ is skew-symmetric in the L^2-inner product with respect to the boundary terms, when using partial integration, and $R(z)$ is symmetric positive semi-definite. To utilize Galerkin projection for space discretization and model reduction, we set up a weak pH-formulation for (2), using the finite element method, see [3]. The following theorem can then be proved analogously to [4].

Theorem 1 *Let $z = [\rho, m, e]^T$ be a strong solution of system (2). Then the following energy dissipation and mass conservation properties hold,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(z) \leq \int_0^L \frac{4k_\omega}{d}T_\infty\mathrm{d}x - \left[m\left(\frac{m^2}{2\rho^2} + \frac{1}{\rho}(e+p)\right)\right]\Big|_0^L \quad \textit{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\int_0^L \rho\mathrm{d}x = -[m]|_0^L\,.$$

To keep these properties on all levels of approximation, we apply Galerkin approximation schemes with compatible finite element bases, such that the pH-structure is preserved. We particularly extend the ansatz from [4], see [3]. Thus, the approximation spaces need to fulfill the following assumption.

Assumption 1 (Compatibility Conditions) *Let* $\mathcal{V} = \mathcal{V}_{\mathsf{\rho}} \times \mathcal{V}_{\mathsf{m}} \times \mathcal{V}_{\mathsf{e}}$ *and* $\mathcal{V}_{\mathsf{\rho}} \subset L^2(\omega)$, $\mathcal{V}_{\mathsf{m}}, \mathcal{V}_{\mathsf{e}} \subset H^1(\omega)$ *be finite dimensional subspaces which fulfill the following assumptions:*

1. $\mathcal{V}_{\mathsf{\rho}} = \partial_x \mathcal{V}_{\mathsf{m}}$ *with* $\partial_x \mathcal{V}_{\mathsf{m}} = \{\xi : \textit{It exists } \zeta \textit{ with } \partial_x \zeta = \xi\}$
2. $\{b \in H^1(\omega) : \partial_x b = 0\} \subset \mathcal{V}_{\mathsf{m}}$
3. $1 \in \mathcal{V}_{\mathsf{e}}$

Let $T_{\mathsf{h}}(\omega) = \{\omega_j\}$ be a uniform partition of $\omega = \cup_{j=1}^n \omega_j = \cup_{j=1}^n [x_j, x_{j+1}]$, where x_i, $i = 1, \dots, n+1$, are the grid points. The space of piecewise polynomials of degree $\leq l$ is defined as $P_l(T_{\mathsf{h}}(\omega)) = \{\varphi \in L^2(\omega) : \varphi|_{\omega_j} \in P_l(\omega_j), \omega_j \in T_{\mathsf{h}}(\omega)\}$. Bases fulfilling Assumption 1 are then, e.g., $\text{span}\{\psi_1, \dots, \psi_n\} = P_0$ for ρ and $\text{span}\{\phi_1, \dots, \phi_{n+1}\} = P_1$ for m and e.

The semi-discretization of system (2) is then given as,

$$\underbrace{\begin{pmatrix} \mathbf{M}_{\mathsf{\rho}} & & \\ \mathbf{M}_{\mathsf{m,\rho}} & \mathbf{M}_{\mathsf{m,m}} & \\ & & \mathbf{M}_{\mathsf{e}} \end{pmatrix}}_{\mathbf{E}(\rho_{\mathsf{h}}, m_{\mathsf{h}})} \underbrace{\begin{pmatrix} \dot{\boldsymbol{\rho}} \\ \dot{\mathbf{m}} \\ \dot{\mathbf{e}} \end{pmatrix}}_{\dot{\mathbf{z}}} = \underbrace{\begin{pmatrix} & \mathbf{J}_{\mathsf{\rho,m}} & \\ -\mathbf{J}_{\mathsf{\rho,m}}^T & & \mathbf{J}_{\mathsf{m,e}} - \tilde{\mathbf{J}}_{\mathsf{m,e}} \\ & -\mathbf{J}_{\mathsf{m,e}}^T + \tilde{\mathbf{J}}_{\mathsf{m,e}}^T & -\mathbf{R}_{\mathsf{e,e}} \end{pmatrix}}_{\mathbf{J}(\rho_{\mathsf{h}}, m_{\mathsf{h}}, e_{\mathsf{h}}) - \mathbf{R}(\rho_{\mathsf{h}}, e_{\mathsf{h}})} \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{m} \\ \mathbf{1} \end{pmatrix}}_{\tilde{\mathbf{e}}(\rho_{\mathsf{h}}, m_{\mathsf{h}})} + \begin{pmatrix} \\ \mathbf{b}_{\mathsf{m}} \\ \mathbf{g}_{\mathsf{e}} + \mathbf{b}_{\mathsf{e}} \end{pmatrix},$$

with $j, q = 1, \dots, n$ and $i, \iota = 1, \dots, n+1$, and

$$\begin{aligned}
&[\mathbf{M}_{\mathsf{\rho}}]_{j,q} = (\psi_q, \psi_j), && [\mathbf{J}_{\mathsf{\rho,m}}]_{j,\iota} = -(\partial_x \phi_\iota, \psi_j), \\
&[\mathbf{M}_{\mathsf{m,\rho}}]_{i,j} = -(\tfrac{m_{\mathsf{h}}}{\rho_{\mathsf{h}}^2} \psi_j, \phi_i), && [\mathbf{J}_{\mathsf{m,e}}]_{i,\iota} = (-\tfrac{e_{\mathsf{h}}}{\rho_{\mathsf{h}}} \partial_x \phi_\iota - \tfrac{1}{\rho_{\mathsf{h}}} \partial_x (p_{\mathsf{h}} \phi_\iota), \phi_i), \\
&[\mathbf{M}_{\mathsf{m,m}}]_{i,\iota} = (\tfrac{1}{\rho_{\mathsf{h}}} \phi_\iota, \phi_i), && [\tilde{\mathbf{J}}_{\mathsf{m,e}}]_{i,\iota} = (-\tfrac{\lambda}{2d} \tfrac{m_{\mathsf{h}}}{\rho_{\mathsf{h}}^2} \left| m_{\mathsf{h}} \right| \phi_\iota, \phi_i), \\
&[\mathbf{M}_{\mathsf{e}}]_{i,\iota} = (\phi_\iota, \phi_i), && [\mathbf{R}_{\mathsf{e,e}}]_{i,j} = (\tfrac{4k_\omega}{d} T_{\mathsf{h}} \phi_j, \phi_i), \\
&[\mathbf{b}_{\mathsf{m}}]_i = -\left[\tfrac{m_{\mathsf{h}}^2}{2\rho_{\mathsf{h}}^2} \phi_i\right]\Big|_0^L, && [\mathbf{b}_{\mathsf{e}}]_\iota = -\left[\tfrac{m_{\mathsf{h}}(e_{\mathsf{h}} + p_{\mathsf{h}})}{\rho_{\mathsf{h}}} \phi_\iota\right]\Big|_0^L, \\
&[\mathbf{g}_{\mathsf{e}}]_\iota = (\tfrac{4k_\omega}{d} T_\infty, \phi_\iota), && \boldsymbol{\varepsilon} = \mathbf{M}_{\mathsf{\rho}}^{-1} \mathbf{f},\ [\mathbf{f}]_j = (\tfrac{m_{\mathsf{h}}^2}{2\rho_{\mathsf{h}}^2}, \psi_j).
\end{aligned} \tag{4}$$

The Hamiltonian is given by $\mathcal{H}(\mathbf{z}) = \frac{1}{2} \mathbf{m}^T \mathbf{M}_{\mathsf{m,m}} \mathbf{m} + \mathbf{1}^T \mathbf{M}_{\mathsf{e}} \mathbf{e}$.

The subscript h denotes the approximated quantities, e.g., $\rho_{\mathsf{h}} = \sum_{j=1}^n \boldsymbol{\rho}_i \psi_i$. By construction, $\mathbf{J}(\rho_{\mathsf{h}}, m_{\mathsf{h}}, e_{\mathsf{h}})$ and $\mathbf{R}(\rho_{\mathsf{h}}, e_{\mathsf{h}})$ are skew-symmetric and positive semi-definite, respectively. Additionally, we have that $\tilde{\mathbf{J}}_{\mathsf{m,e}} = \tilde{\mathbf{J}}_{\mathsf{m,e}}^T$ and $\mathbf{R}_{\mathsf{e,e}} = \mathbf{R}_{\mathsf{e,e}}^T \geq 0$. Furthermore, $\mathbf{E}(\rho_{\mathsf{h}}, m_{\mathsf{h}})^T \tilde{\mathbf{e}}(\rho_{\mathsf{h}}, m_{\mathsf{h}}) = \nabla_{\mathbf{z}} \mathcal{H}(\mathbf{z})$ can be verified.

3 Structure-Preserving Model Order Reduction

We show how the pH-structure of system (4) is preserved during projection-based model reduction. The reduction relies on the snapshot matrix $\mathbf{S}$, i.e.,

$$\mathbf{S} = (\mathbf{S}_\rho^T\, \mathbf{S}_\mathsf{m}^T\, \mathbf{S}_\mathsf{e}^T)^T,\ \mathbf{S}_\rho = (\boldsymbol{\rho}^0, \dots, \boldsymbol{\rho}^{n_t}),\ \mathbf{S}_\mathsf{m} = (\mathbf{m}^0, \dots, \mathbf{m}^{n_t}),\ \mathbf{S}_\mathsf{e} = (\mathbf{e}^0, \dots, \mathbf{e}^{n_t}). \quad (5)$$

Here, n_t denotes the number of time points used in the simulation. To keep properties like energy dissipation, the reduced spaces $\mathcal{V}_{\rho,\mathsf{r}} \subset \mathcal{V}_\rho$, $\mathcal{V}_{\mathsf{m},\mathsf{r}} \subset \mathcal{V}_\mathsf{m}$ and $\mathcal{V}_{\mathsf{e},\mathsf{r}} \subset \mathcal{V}_\mathsf{e}$ need to fulfill Assumption 1. For this, we take advantage of the algebraic equivalent of Assumption 1, as done in [2], i.e.,

Assumption 2 (Algebraic Compatibility Conditions) *Let the reduction basis* $\mathbf{V}_r$ *have block-diagonal structure, i.e.,* $\mathbf{V}_r = blkdiag(\mathbf{V}_\rho, \mathbf{V}_m, \mathbf{V}_e)$*. Then* $\mathbf{V}_r$ *is assumed to fulfill*

1. $\text{image}(\mathbf{M}_\rho \mathbf{V}_\rho) = \text{image}(\mathbf{J}_{\rho,m} \mathbf{V}_m)$
2. $\text{kernel}(\mathbf{J}_{\rho,m}) \subset \text{image}(\mathbf{V}_m)$
3. $(1, \dots, 1)^T \in \text{image}(\mathbf{V}_e)$

The block-diagonal projection matrix $\mathbf{V}_\mathsf{r}$ additionally preserves the block structure of system (4). To obtain $\mathbf{V}_\rho$, $\mathbf{V}_\mathsf{m}$ and $\mathbf{V}_\mathsf{e}$, we start by computing $\mathbf{V}_\rho$ from the snapshots (5) by Algorithm 1, which uses proper orthogonal decomposition.

Algorithm 1 (Computation of $\mathbf{V}_\rho$) *Let* $\mathbf{S}$, $\mathbf{M}_\rho$, $\mathbf{J}_{\rho,m}$ *and* r_ρ *be given.*

1. Set up the snapshot matrices $\mathbf{S}_\rho$, $\mathbf{S}_m$ *and* $\mathbf{S}_e$ *from* $\mathbf{S}$.
2. Construct a reduced basis $\mathbf{V}_\rho$ *from* $[\mathbf{S}_\rho,\ \mathbf{M}_\rho^{-1}(\mathbf{J}_{\rho,m}[\mathbf{S}_m,\ \mathbf{S}_e])]$ *by POD w.r.t. the scalar product induced by* $\mathbf{M}_\rho$ *of dimension* r_ρ.

In the computation of $\mathbf{V}_\rho$ we do not only make use of $\mathbf{S}_\rho$, but also of $\mathbf{M}_\rho^{-1}\mathbf{J}_{\rho,\mathsf{m}}\mathbf{S}_\mathsf{m}$. This is motivated by Assumption 2-1. It also resembles using the time derivatives of ρ in POD. Additionally, our numerical experiments show that adding $\mathbf{M}_\rho^{-1}\mathbf{J}_{\rho,\mathsf{m}}\mathbf{S}_\mathsf{e}$ to the computation of $\mathbf{V}_\rho$ enhances the approximation quality of the reduced bases. As we construct $\mathcal{V}_\mathsf{m}$ and $\mathcal{V}_\mathsf{e}$ with the same bases, it follows that $\mathbf{S}_\mathsf{e} \in \mathcal{V}_\mathsf{m}$, such that Assumption 2-1 is not violated by including these snapshots. The compatibility conditions are enforced, as seen in Algorithm 2. Here, $\mathbf{V}_\mathsf{m}$ and $\mathbf{V}_\mathsf{e}$ are deduced from $\mathbf{V}_\rho$, which contains information of all dynamic states, i.e., $\boldsymbol{\rho}$, $\mathbf{m}$ and $\mathbf{e}$.

Algorithm 2 (Compatible Basis) *Let* $\mathbf{V}_\rho$, $\mathbf{M}_\rho$, $\mathbf{M}_m$, $\mathbf{J}_{\rho,m}$ *be given.*

1. Compute $\mathbf{N} = \text{kernel}(\mathbf{J}_{\rho,m})$.
2. Compute $\mathbf{V}_m$ *by*
   a. $\mathbf{W}_m = \mathbf{M}_m^{-1}\mathbf{J}_{\rho,m}^T(\mathbf{J}_{\rho,m}\mathbf{M}_m^{-1}\mathbf{J}_{\rho,m}^T)^{-1}\mathbf{M}_\rho\mathbf{V}_\rho$,
   b. Orthogonalize $\mathbf{N}$ *to columns of* $\mathbf{W}_m$ *w.r.t. inner product induced by* $\mathbf{M}_m$.
3. Set $\mathbf{V}_e$ *to be equal to* $\mathbf{V}_m$.

Here, $\mathbf{M}_\mathsf{m}$ denotes the mass matrix related to $\mathcal{V}_\mathsf{m}$. Lines (2a) and (2b) in Algorithm 2 enforce the Assumptions 2-1 and 2-2 for the space $\mathcal{V}_{\mathsf{m},\mathsf{r}}$. Choosing the same basis for $\mathcal{V}_{\mathsf{m},\mathsf{r}}$ and $\mathcal{V}_{\mathsf{e},\mathsf{r}}$ might seem odd, but experiments show that this choice enhances the stability of the reduced models. It creates more symmetry in the reduced models, as $\tilde{\mathbf{J}}_{\mathsf{m},\mathsf{e}}$ and $\tilde{\mathbf{J}}_{\mathsf{e},\mathsf{m}}$ are still symmetric after reduction. Furthermore, the Assumption 2-3 is fulfilled. This follows from $\operatorname{span}\{\mathbf{1}\} = \operatorname{span}\{\operatorname{kernel}(\mathbf{J}_{\rho,\mathsf{m}})\}$, as the kernel of $\mathbf{J}_{\rho,\mathsf{m}}$ is the kernel of the discretized gradient operator. Finally, we compute the reduced order model by multiplying (4) with $\mathbf{V}_\mathsf{r}^T$ from the left and by multiplying $\mathbf{E}(\rho_\mathsf{h}, m_\mathsf{h})$, $\mathbf{J}(\rho_\mathsf{h}, m_\mathsf{h}, e_\mathsf{h})$ and $\mathbf{R}(\rho_\mathsf{h}, e_\mathsf{h})$ by $\mathbf{V}_\mathsf{r}$ from the right. The reduced efforts are given by $\varepsilon_\mathsf{r} = \mathbf{V}_\rho^\dagger \varepsilon$ and $\mathbf{1}_\mathsf{r} = \mathbf{V}_\mathsf{e}^\dagger \mathbf{1}$, with $\mathbf{V}_\alpha^\dagger = (\mathbf{V}_\alpha^T \mathbf{M}_\alpha \mathbf{V}_\alpha)^{-1} \mathbf{V}_\alpha^T \mathbf{M}_\alpha$, $\alpha \in \{\rho, \mathsf{m}, \mathsf{e}\}$ being the corresponding pseudo-inverse.

4 Numerical Results and Discussion

Considering the flow of an ideal compressible gas through a single pipe in the subsonic regime, we have $p = \frac{R}{c_V} e$ and $T = \frac{1}{c_V} \frac{e}{\rho}$. The parameters are given by $R = 1$, $c_V = \frac{5}{2}$, $L = 1$, $d = 1$, $\lambda = 40$, $k_\omega = \frac{5}{4}$, $T_\infty = 1$, similar to the second example in [1]. The initial values at $t = 0$ and the boundary conditions at $x = 0$ are chosen to be

$$\rho(0, x) = 3,\ m(0, x) = 0.3,\ e(0, x) = 9 \quad \text{and} \quad m(t, 0) = m(t, 1) = 0.3,\ e(t, 0) = 9,$$

respectively. The full order model (FOM) is discretized with a spatial step size $\Delta x = 0.01$, resulting in a dimension of $3n + 2 + 3 = 305$ with $n = 100$. The latter $+3$ results from the Lagrange multipliers, which are used to couple the boundary conditions to the system and consistently initialized. The Lagrange multipliers will not be reduced. The FOM is simulated using the implicit Euler scheme with $\Delta t = 0.1$ and Newton's method with an analytical Jacobian until the stationary solution is reached at $t_f = 30s$. The reduced models (ROMs) are solved using the same boundary conditions and parameters as the FOM. With the snapshot matrix (5), set up from the solutions of the FOM, we compute the projection matrices with Algorithms 1 and 2. The reduced initial values are given as,

$$\rho_\mathsf{r}(0) = \mathbf{V}_\mathsf{r}^\dagger \rho(0), \quad \mathbf{m}_\mathsf{r}(0) = \mathbf{V}_\mathsf{m}^\dagger \mathbf{m}(0), \quad \mathbf{e}_\mathsf{r}(0) = \mathbf{V}_\mathsf{e}^\dagger \mathbf{e}(0).$$

Numerical tests show that the ROMs computed from our pH-FOM are only stable, if the parity of r_ρ is equal to that of n. This might be explained by the different constitutions of the kernel of the skew-symmetric system matrix $\mathbf{J}(\rho_\mathsf{h}, m_\mathsf{h}, e_\mathsf{h})$ for even or odd numbers of degrees of freedom. This question is still part of ongoing research. As $n = 100$, we only consider ROMs with r_ρ even as well. Figure 1 shows the maximum relative L^2-errors for reduction E_t and projection $E_{t,P}$ for

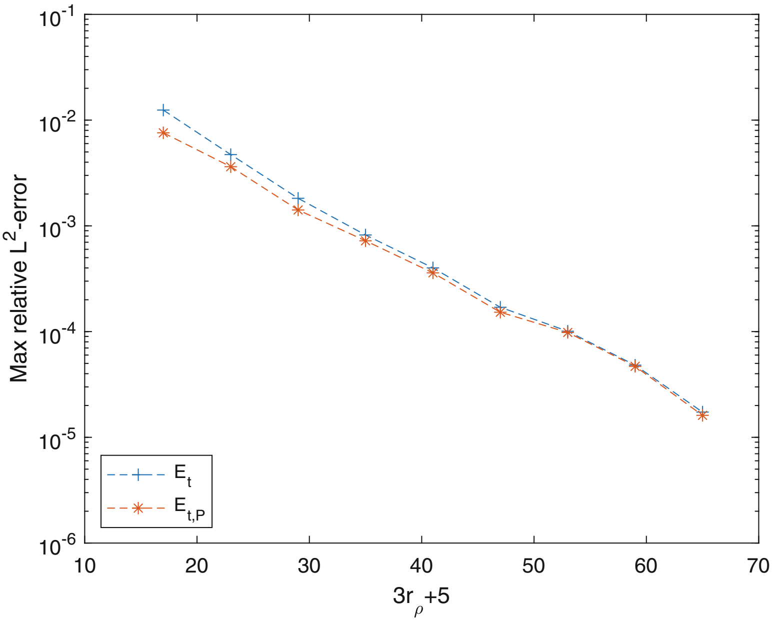


Fig. 1 Reduction and projection errors over r_ρ for z

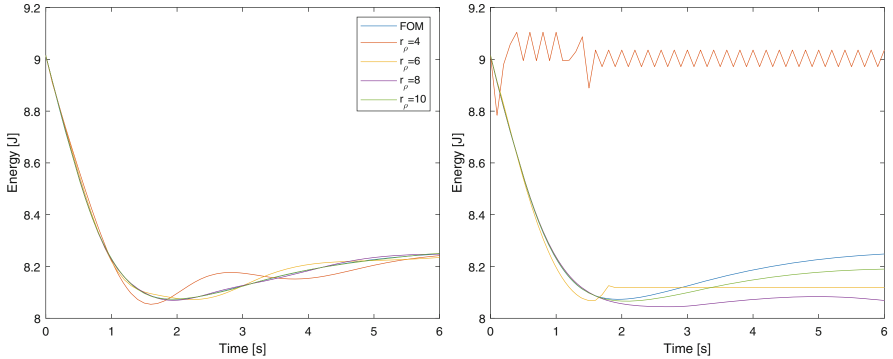


Fig. 2 Total energy with and without compatibility conditions

different even r_ρ, i.e., $\max_{t\in[0,t_f]} \frac{||z(t,x)-z_{\mathbf{r}}(t,x)||_{L^2}}{||z(t,x)||_{L^2}}$. With increasing dimension the errors decline rapidly. This excellent error behavior is only due to the compatibility conditions from Assumption 2, see Figs. 2 and 3. When they are not used to compute the projection bases, the ROMs do not converge or loose energy dissipation and mass conservation, for the r_ρ, where the systems fulfilling Assumption 2 already converge or fulfill the properties. Summing up, the numerical example yields

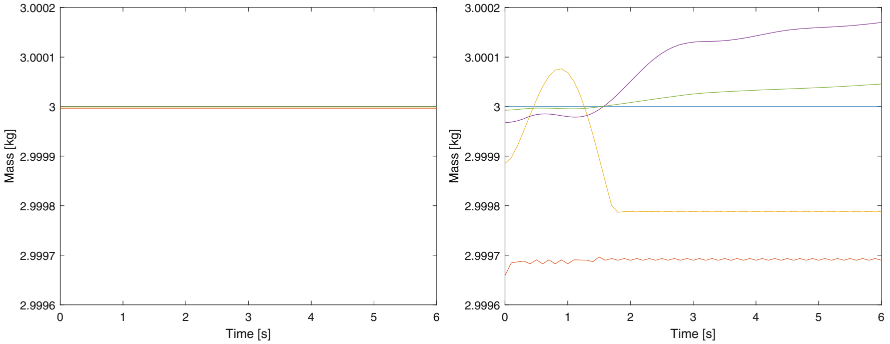


Fig. 3 Total mass with and without compatibility conditions

promising results for structure-preserving reduction of a pipe flow. The extension to networks and complexity reduction are part of current research.

Acknowledgments The authors acknowledge the support of the BMBF, Project EiFer.

References

1. H. Egger. A mixed variational discretization for non-isothermal compressible flow in pipelines. arXiv:1611.03368, 2016.
2. H. Egger, T. Kugler, B. Liljegren-Sailer, N. Marheineke, and V. Mehrmann. On structure-preserving model reduction for damped wave propagation in transport networks. SIAM J. Sci. Comput., 40(1), 2018.
3. S.-A. Hauschild and N. Marheineke. Structure-preserving discretization of a port-Hamiltonian formulation of the non-isothermal Euler equations. PAMM, 20(1):e202000014, 2021.
4. B. Liljegren-Sailer and N. Marheineke. On port-Hamiltonian approximation of a nonlinear flow problem on networks. arXiv:2009.11216, 2021.
5. V. Mehrmann and R. Morandin. Structure-preserving discretization for port-Hamiltonian descriptor systems. In: 58th IEEE Conference on Decision and Control, CDC 2019, Nice, France, December 11–13, 2019, pages 6863–6868. IEEE, 2019.

ECMI Modelling Week: First Time in Russia and First Time Online

Tatiana Pogarskaia, Sergey Lupuleac, and Matti Heiliö

Abstract The European Consortium for Mathematics in Industry (ECMI) has been running annual Modelling Weeks (MW) for students since 1988. Students come from all over Europe to spend a week working in small multinational groups on real-life based problems. While the COVID-19 pandemic educational process was changed to online format in the majority of universities all over the world in order to continue teaching with no interruption. The Modeling Week 2020 hosted by SPbPU was held online as well but kept it as close to traditional ECMI Modelling Week as possible.

The aim of the paper is to compare the performance of the participants who attended the first Online Modelling Week in distance learning format with previous ones held in a face-to-face format. The example of 30 students from different universities and countries divided into 4 groups is considered. The group work analysis revealed that more students dropped out of the Modelling Week in comparison to previous years. At the same time the smaller groups were able to solve the problems and finish the course.

1 Introduction

Technological advances provided an opportunity to share knowledge in the online format what made it very attractive. The problem of the effectiveness of distance and online learning methodologies has been discussed in recent decades before the COVID-19 pandemic [1–3]. Among other things, comparison of online and

T. Pogarskaia (✉) · S. Lupuleac
Institute of Applied Mathematics and Mechanics, Peter the Great St.Petersburg Polytechnic University, St Petersburg, Russia
e-mail: pogarskaya_ta@spbstu.ru

M. Heiliö
Computational and Process Engineering, LUT University, Lappeenranta, Finland
e-mail: Matti.Heilio@lut.fi

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_2

in-class education systems was regarded from different aspects [4–6], including specificity of online teaching technical subjects [7, 8]. Undoubtedly, distance learning during the pandemic differs from the previous experience primarily by the social restrictions and the psychological state of both students and teachers [9–12]. It makes the education process to be a great challenge as it requires adapting existing programs and methodologies.

One of the main distinctive features of the ECMI MW is its format implying a week working in small multinational groups on projects which are based on real-life problems. Due to the short time period and the need to solve a problem together, students acquire skills of teamwork and communication. We tried to keep as close to traditional ECMI MW as possible and save the average group size (5–7 participants), their age (senior bachelor students and master students) and projects origins. The purpose of this article is to compare the results of participants who attended the first Online Modeling Week in the distance format with the results of previously face-to-face events.

2 The First Online Modelling Week—New Solution

The Modelling Week was going to be held in Peter the Great St.Petersburg Polytechnic University (SPbPU), St.Petersburg, Russia, in July (05.07.2020–12.07.2020) [13]. As it was mentioned above, the event can be described as following. Each group of usually 5–6 students is led by an ECMI instructor who introduces the problem formulated in non-mathematical terms and helps to guide the students to a solution during the week. The students present their results to the other groups on the last day and then write up their work as a report. The preparations were started in March 2019 but the pandemic changed the situation and the format was changed to online. The first problem we faced was connected with registration declines as 6 of 8 instructors and 9 of 13 students registered before the pandemic refused to take part in the new format.

However, four different projects were proposed by instructors from Wroclaw University of Science and Technology, Leeds Beckett University and SPbPU and 30 students applied to participate. The needed mathematical background was nearly the same for all the problems and included calculus and optimization methods.

SPbPU has been providing online courses since 2015 and the experience was widely developed during the pandemic and was used during the MW. MS Teams platform was used for the group work as it allows a wide range of opportunities such as video call, file sharing, chatting etc. The projects were announced in advance and the students were assigned to the problem of their choice after an e-mail survey. The instructors acted as coordinators and played minimal roles unless students needed some more explanations.

3 Results

The day before the MW started 4 students informed that they declined their participation and finally there were 7, 6, 6 and 7 students in the groups of projects 1, 2, 3, and 4, respectively. Projects 1 and 4 included the main and the associated instructors. The academic level of the students is presented in Table 1.

The first project called "Bolted assembly optimization" was aimed at the loosening of bolts phenomenon in aircraft assembly. Students were proposed to study it by using the special assembly demonstrator developed on the base of a specialized software for simulation of the aircraft assembly process [14, 15]. The second project, "Hybrid Storage System", proposed to construct and emulate the performance of the existing algorithm for two consecutive days on the real datasets employing REDWoLF mathematical template [16]. The aim of the next project, "Capillary moisture uptake", was to deliver a mathematical model of the capillary moisture absorption in wood and to perform the analysis of the distribution of water content in the trunk. Since the absorption of water may differ for different types of trees thus the investigation of various types of wood was needed. And the last project was based on the short Science Fiction novel "Jack and the Beanstalk" by Richard A. Lovett about a guy who climbs a tower that is 65000 kilometres high. It was proposed to make technical audit of the project starting with a simple model: Earth as massive ball, long infinitely narrow and rigid beam with distributed mass (the Beanstalk) and space shuttle (massive dot) to be launched to Mars.

The participants represented universities from Germany, Serbia, Italy, Portugal and China. Due to different time zones in the range from Portugal to China, the groups were allowed to choose the most appropriate time for work, time 11:00–16:00 (GMT+3) was recommended. The duration of the work remained approximately the same but it was decreased from 5 and a half to 5 days as the first day traditionally spent for arrivals was not needed. The program implied a half working Monday with problem presentations in the morning, full days from Tuesday to Friday and closing day on Saturday with Zoom final presentations and discussion. Figure 1 represents the distribution of the worked time by days during the MW for each group. It can be noticed that groups 2 and 3 with the less number of students worked more steadily and groups 1 and 4 spent much more time on their work on the last day before they had to present the results.

The groups were free to choose the way to communicate and all of them used video calls but time could vary. The percentage of worked time spent on video calls is presented in Fig. 2. The data were collected via MS Teams. It must be mentioned that group 4 several times used Zoom platform and the time was not evaluated.

Table 1 Students academic level over the teams

Academic degree	Project 1	Project 2	Project 3	Project 4
Bachelor students	4	1	3	4
Master students	3	5	3	2
PhD	0	0	0	1

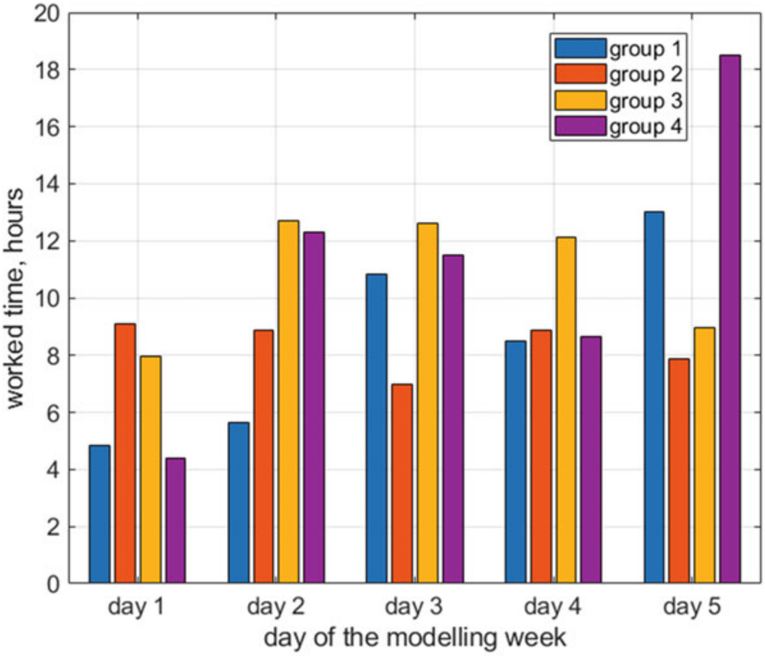


Fig. 1 The distribution of the worked time by each day during the MW for each group, hours

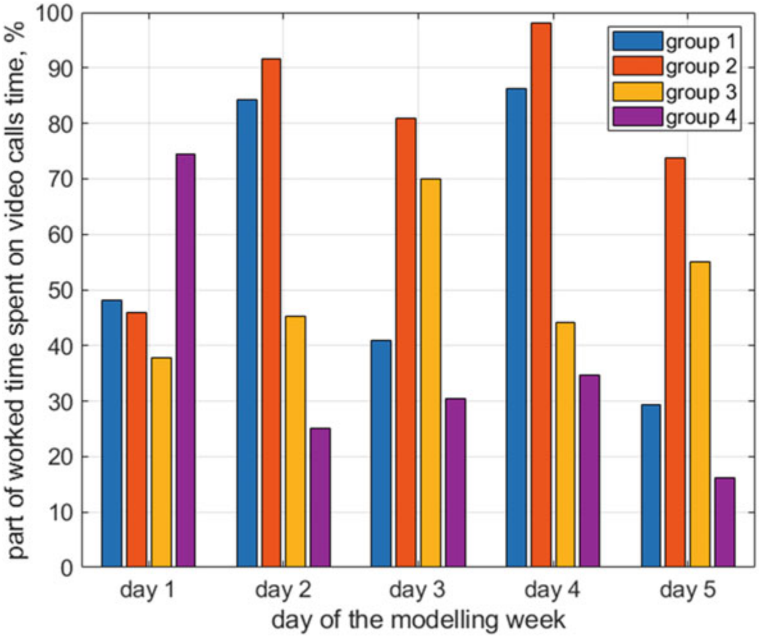


Fig. 2 The time the groups spent among themselves on video calls for discussions

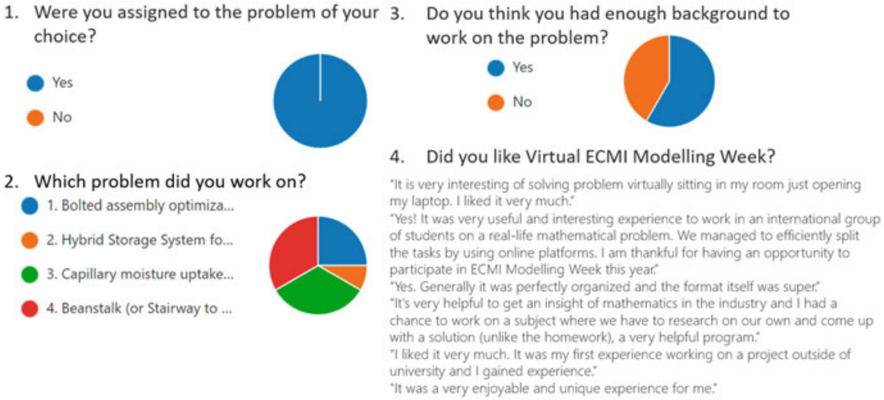
Only group 2 spent nearly all their time in touch via video. Groups 1 and 3 spent on average the majority of the total time communicating via video as well.

All the groups managed to finish their work on time and prepare presentations. The next step was to prepare a report summarizing their work until the end of October to get 3 ECTS credits and certificates. Students of the second project published the results of their work [17]. Finally, only 20 students (of 30 registrated) completed the task. In comparison to the previous ECMI Modelling Weeks which were held in a face-to-face format, the part of students who did not complete the course is very high. For example, 32nd Modelling Week in Novi Sad on 2018 was successfully finished by all the participants; 29th Modelling Week held in Lisbon in 2015 and ECMI Summer School and Modelling Week in Milan in 2010 had 1 and 2 drop-out students of 55 and 85 respectively [18].

4 Conclusion

Online learning was regarded as the future educational format when students are given more flexibility even several years ago. The COVID-19 pandemic showed the irreplaceable help of distance learning format to continue education with minimal interruptions that cannot be overemphasized.

At the same time, we can notice that it was challenging for instructors to engage students in work. We asked the instructors to comment on their experience and all of them mentioned that it was hard to explain the task online and only students with the best background were fully involved. The high part of students (33%) who did not complete the course or decided not to take part in the change to online format can additionally illustrate the weaknesses of virtual format such as reduction of student engagement and loss of assessments that were noticed in [19, 20]. As a possible solution, we regard a reward of the best group in the form of an invitation to take part in a conference or publication of the project results. However, a short survey the majority of students finished provided us with only positive feedback (Fig. 3).



1. Were you assigned to the problem of your choice?

2. Which problem did you work on?

● 1. Bolted assembly optimiza...
● 2. Hybrid Storage System fo...
● 3. Capillary moisture uptake...
● 4. Beanstalk (or Stairway to ...

3. Do you think you had enough background to work on the problem?

4. Did you like Virtual ECMI Modelling Week?

"It is very interesting of solving problem virtually sitting in my room just opening my laptop. I liked it very much."
"Yes! It was very useful and interesting experience to work in an international group of students on a real-life mathematical problem. We managed to efficiently split the tasks by using online platforms. I am thankful for having an opportunity to participate in ECMI Modelling Week this year."
"Yes. Generally it was perfectly organized and the format itself was super."
"It's very helpful to get an insight of mathematics in the industry and I had a chance to work on a subject where we have to research on our own and come up with a solution (unlike the homework), a very helpful program."
"I liked it very much. It was my first experience working on a project outside of university and I gained experience."
"It was a very enjoyable and unique experience for me."

Fig. 3 The survey of the students after the end of the MW

Acknowledgments The first Online Modelling Week was supported by the Academic Excellence Project 5–100 proposed by Peter the Great St. Petersburg Polytechnic University.

References

1. R.M. Bernard, P.C. Abrami, Y. Lou, E. Borokhovski, A. Wade, L. Wozney, P.A. Wallet, M. Fiset, and B. Huang, How does distance education compare with classroom instruction? A meta-analysis of the empirical literature, Review of Educational Research, vol. 74, no. 3, pp. 379–439, 2004.
2. J. O'Malley and H. McCraw, Students perceptions of distance learning, online learning and the traditional classroom, Online Journal of Distance Learning Administration, vol. 2, 1999.
3. M. Shachar and Y. Neumann, Differences between traditional and distance education academic performances: A meta-analytic approach, The International Review of Research in Open and Distributed Learning, vol. 4, Oct. 2003.
4. S.R. Sankaran, D. Sankaran, and T. Bui, Effect of student attitude to course format on learning performance: An empirical study in web vs. lecture instruction, Journal of Instructional Psychology, vol. 27, p. 66, 2000.
5. R. Schoenfeld-Tacher, S. McConnell, and M. Graham, Do no harm – a comparison of the effects of on-line vs. traditional delivery media on a science course, Journal of Science Education and Technology, vol. 10, no. 3, pp. 257–265, 2001.
6. M.D.B. Castro and G.M. Tumibay, A literature review: efficacy of online learning courses for higher education institution using meta-analysis, Education and Information Technologies, vol. 26, no. 2, pp. 1367–1385, 2021.
7. L.M. O'Dwyer, R. Carey, and G. Kleiman, A study of the effectiveness of the Louisiana algebra I online course, Journal of Research on Technology in Education, vol. 39, no. 3, pp. 289–306, 2007.
8. J. Kleinman and E.B. Entin, Comparison of in-class and distance-learning students' performance and attitudes in an introductory computer science course, J. Comput. Sci. Coll., vol. 17, p. 206–219, May 2002.
9. R.C. Chick, G.T. Clifton, K.M. Peace, B.W. Propper, D.F. Hale, A.A. Alseidi, and T.J. Vreeland, Using technology to maintain the education of residents during the COVID-19 pandemic, Journal of Surgical Education, vol. 77, no. 4, pp. 729–732, 2020.
10. C. Foo, B. Cheung, and K. Chu, A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic, BMC Medical Education, vol. 21, no. 1, 2021.
11. S.I. Hofer, N. Nistor, and C. Scheibenzuber, Online teaching and learning in higher education: Lessons learned in crisis situations, Computers in Human Behavior, vol. 121, 2021.
12. M. Bonati, R. Campi, M. Zanetti, M. Cartabia, F. Scarpellini, A. Clavenna, and G. Segre, Psychological distress among italians during the 2019 coronavirus disease (COVID-19) quarantine, BMC Psychiatry, vol. 21, no. 1, 2021.
13. "Virtual ECMI modelling week." http://mw2020.spbstu.ru/.
14. S. Lupuleac, A. Smirnov, J. Shinder, M. Petukhova, M. Churilova, E. Victorov, and J. Bouriquet, Complex fastener model for simulation of airframe assembly process, in: ASME International Mechanical Engineering Congress and Exposition, Proceedings (IMECE), vol. 2B-2020, 2020.
15. S. Lupuleac, T. Pogarskaia, M. Churilova, M. Kokkolaras, and E. Bonhomme, Optimization of fastener pattern in airframe assembly, Assembly Automation, vol. 40, no. 3, pp. 723–733, 2020.
16. A.A. Shukhobodskiy and G. Colantuono, Red wolf: Combining a battery and thermal energy reservoirs as a hybrid storage system, Applied Energy, vol. 274, 2020.

17. M. Wiesheu, L. Rutešić, A.A. Shukhobodskiy, T. Pogarskaia, A. Zaitcev, and G. Colantuono, Red wolf hybrid storage system: Adaptation of algorithm and analysis of performance in residential dwellings, Renewable Energy ,2021.
18. Modelling weeks, https://ecmiindmath.org/education/modelling-weeks/.
19. R. Wilcha, Effectiveness of virtual medical teaching during the COVID-19 crisis: Systematic review, JMIR Medical Education, vol. 6, no. 2, 2020.
20. S. Kaup, R. Jain, S. Shivalli, S. Pandey, and S. Kaup, Sustaining academics during COVID-19 pandemic: The role of online teaching-learning, Indian Journal of Ophthalmology, vol. 68, no. 6, pp. 1220–1221, 2020.

Parameter Calibration with Consensus-Based Optimization for Interaction Dynamics Driven by Neural Networks

Claudia Totzeck and Simone Göttlich

Abstract We calibrate parameters of neural networks that model forces in interaction dynamics with the help of the Consensus-based global optimization method (CBO). We state the general framework of interaction particle systems driven by neural networks and test the proposed method with a real dataset from the ESIMAS traffic experiment. The resulting forces are compared to well-known physical interaction forces. Moreover, we compare the performance of the proposed calibration process to the one in Göttlich and C. Totzeck (Optimal control for interaction particle systems driven by neural networks. arXiv:2101.12657, 2021) which uses a stochastic gradient descent algorithm.

1 Introduction

Modelling interacting particle dynamics such as traffic, crowd dynamics, schools of fish and flocks of birds has attracted the attention of many research groups in the recent decades. Most models use physically-inspired interaction forces resulting from potentials to capture the observed behaviour. In fact, the gradient of the potential is used as driving force for interacting particle systems formulated with the help of ordinary differential equation (ODE). These models are able to represent the main features of the dynamics, but as for all models we cannot be sure that they deliver the whole truth. The idea in [1] was therefore to replace the physical-inspired models by neural networks, train the networks with real data and compare the resulting forces.

C. Totzeck (✉)
School of Mathematics and Natural Sciences, University of Wuppertal, Wuppertal, Germany
e-mail: totzeck@uni-wuppertal.de

S. Göttlich
School of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany
e-mail: goettlich@uni-mannheim.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_3

In the recent years it became obvious that neural networks are able to represent a lot of details from the dataset. It may be possible that there are details captured that are not even noticed by humans and therefore do not appear in physical models which are built to reproduce observations of the modeller.

In the following we recall the general dynamic of interaction particle systems driven by neural networks as proposed in [1]. Then we shortly describe the global optimization method 'Consensus-based optimization' that we use for the real-data based calibration the network. Finally, we present the numerical results obtained by the calibration process and compare them to the ones resulting from the calibration with the stochastic gradient descent method reported in [1].

2 Interacting Particle Systems Driven by Neural Networks

We consider interacting particle dynamics described by ODE systems of the form

$$\frac{d}{dt}y_i = \sum_{j=1}^{N} W_\theta^{i,j}(y_j - y_i), \quad y_i(0) = z_0^i, \quad i = 1, \dots, N, \tag{1}$$

where $W_\theta^{i,j}$ represents the interaction force resulting for y_i in its interaction with y_j. The initial condition of the particles is given by real dataset $z_0 = z(0)$. In order to compare the results to the ones in [1] we restrict the class of neural networks to feed-forward networks. However, note that the approach discussed here allows for general neural networks while the discussion in [1] considers feed-forward networks and can only be generalized to neural networks allowing for back propagation.

2.1 Feed-Forward Neural Networks

In the following we consider feed-forward artificial neural networks of the form

Definition 1 A *feed-forward artificial neural network (NN)* is characterized by

- Input layer:

$$a_1^{(1)} = 1, \quad a_k^{(1)} = x_{k-1}, \quad \text{for } k \in \{2, \dots, n(1) + 1\},$$

  where $x \in \mathbb{R}^{n^{(1)}}$ is the input (feature) in (1) and $n^{(1)}$ is the number of neurons without the bias unit a_1.
- Hidden layers: for $\ell \in \{2, \dots, L-1\}, k \in \{2, \dots, n^{(\ell)} + 1\}$

$$a_1^{(\ell)} = 1, \quad a_k^{(\ell)} = g^{(\ell)}\left(\sum_{j=1}^{n^{(\ell-1)}+1} \theta_{j,k}^{(\ell-1)} a_j^{(\ell-1)}\right).$$

– Output layer: $a_k^{(L)} = g^{(L)}\left(\sum_{j=1}^{n^{(L-1)}+1} \theta_{j,k}^{(L-1)} a_j^{(L-1)}\right)$ for $k \in \{1, \dots, n^{(L)}\}$

Note that the output layer has no bias unit. The entry $\theta_{j,k}^{\ell}$ of the weight matrix $\theta^{(\ell)} \in \mathbb{R}^{n^{(\ell-1)} \times n^{(\ell)}}$ describes the weight from neuron $a_j^{(\ell-1)}$ to the neuron $a_k^{(\ell)}$. For notational convenience, we assemble all entries $\theta_{j,k}^{(\ell)}$ in a vector $\mathbb{R}^K$ with

$$K := n^{(1)} \cdot n^{(2)} + n^{(2)} \cdot n^{(3)} + \cdots + n^{(L-1)} \cdot n^{(L)}.$$

For the numerical experiment we use $g^{(\ell)} = \log(1 + e^x)$ for $\ell = 2, \dots, N - 1$ and $g^{(L)}(x) = x$. For an illustration of the NN structure we refer to [1]. In the numerical section we consider an NN with $L = 3$, one input and 5 units in the hidden layer.

3 Parameter Calibration

We formulate the task of the parameter calibration as an optimization problem. Let $u \in \mathbb{R}^d$ denote the vector of parameters to be calibrated. This could be the weights of the neural network θ and some other parameters, as for example the average length L and the maximal speed $v_{\max}$ of the cars which we will consider in the application. As we want the network to recover the forces hidden in the real data dynamics, we define the cost function for the parameter calibration as

$$J(y, u) = \frac{1}{2} \int_0^T \|y(t) - z(t)\|^2 dt + \frac{\delta}{2} |u - u_{\text{ref}}|^2, \tag{2}$$

where z denotes the trajectories of the cars obtained by the traffic experiment, and u_{ref} are reference values for the parameters. The parameter δ allows to balance the two terms in the cost functional. In case no reference values of the parameters are available, we set $\delta = 0$ in the numerical section.

3.1 Consensus-Based Optimization (CBO)

We solve the parameter calibration problem with the help of a Consensus-based optimization method [4] and choose the variant introduced in [5] which is tailored for high-dimensional problems involving the calibration of neural networks. The CBO dynamics is itself a stochastic interacting particle system with N_{CBO} agents given by stochastic differential equations (SDEs). The evolution of the agents is influenced by two terms. On the one hand, there is a deterministic term that aims to confine the positions of the agents at a weighted mean. On the other hand, there is a stochastic term that allows for exploration of the state space. The details are

$$du_t^i = -\lambda(u_t^i - v_f)dt + \sigma \text{diag}(u_t^i - v_f)dB_t^i, \quad i = 1, \dots, N_{\text{CBO}} \tag{3}$$

with drift and diffusion parameters $\lambda, \sigma > 0$, independent d-dimensional Brownian motions B_t^i and initial conditions u_0^i drawn uniformly from the parameter set of interest. A main role plays the weighed mean

$$v_f = \frac{1}{\sum_{i=1}^{N_{\text{CBO}}} e^{-J(u_i)}} \sum_{i=1}^{N_{\text{CBO}}} u_i \, e^{-\alpha J(u_i)}.$$

By its construction, agents with lower cost have more weight in the mean as the ones with higher cost. The parameter α allows to adjust this difference of the weights. For more information on the CBO method and its proof of convergence on the mean-field level we refer the interested reader to [2] and the references therein. As indicated by the notation above, the agents used in the CBO method are different realizations of parameter vectors that we consider for the calibration. For the numerical results NN4 we consider a neural network with 13 weights, i.e., $\theta \in \mathbb{R}^{13}$. Moreover, we assume the maximal speed $v_{\max}$ as additional parameter. Hence, for fixed t we have for the i-th CBO agent $u_t^i \in \mathbb{R}^{14}$.

4 Numerical Results and Conclusion

For the calibration of the parameters we consider real data from the project ESIMAS [3]. As we want to compare the results to the well-known follow-the-leader model for traffic flow (LWR) we recall its details

$$\frac{d}{dt} y_i(t) = f\left(\frac{y_{i+1}(t) - y_i(t)}{L}\right), \quad i = 1, \dots, N-1, \tag{4a}$$

$$\frac{d}{dt} y_N(t) = v_{\max}. \tag{4b}$$

Here $f(\cdot)$ is either $v_{\max} \log(\cdot)$ or $v_{\max}(1 - 1/\cdot)$. To be prepared for a reasonable comparison, we consider for the neural network dynamics

$$\frac{d}{dt} y_i(t) = W_\theta^{i,i+1}(y_{i+1}(t) - y_j(t)), \quad i = 1, \dots, N-1, \tag{5a}$$

$$\frac{d}{dt} y_N(t) = v_{\max} \tag{5b}$$

supplemented with initial data $y(0) = z_0$. This leads to $u = (v_{\max}, \theta)$. To evaluate the models and compute the corresponding cost we solve all ODEs with an explicit Euler scheme. For details we refer to [1]. The number in the notation $NN2$, $NN4$ and $NN10$ corresponds to the number of nonbias neurons in the hidden layer.

4.1 Data Processing and Numerical Schemes

The data collection of the ESIMAS project contains vehicle data from 5 cameras that were placed in a $1km$ tunnel section on the German motorway A3 nearby Frankfurt/Main [3]. The data is processed in the exact same way as in [1]. Files with the processed data can be found online.[1]

The SDE which represents the CBO scheme is solved with the scheme proposed in [5]. In particular, we set $dt = 0.05$, $\sigma_0 = 1$, $\lambda = 1$ and the maximal number of time steps to 100. The mini-batch size of the CBO scheme is 50 and we have 100 CBO agents in total. In each time step we update one randomly chosen mini-batch. The initial values are chosen as follows

$$v_{\max} \sim U([20, 40]), \quad L \sim U([0, 10]) \text{ and } \theta \sim U([-0.5, 0.5]^K)$$

4.2 Resulting Forces and Comparison

Figure 1 (left) shows the velocities resulting from the parameter calibration process. We find that the estimated velocities for the NN approaches are higher than the velocities of the LWR based models. The difference is most significant in data set 10. The plot on the right shows the average of the resulting forces for the different models. The forces of the NN approaches resemble linear approximations of the forces corresponding to the LWR models. The car length (L) appears only in the LWR models. Its optimized values for the different data sets are given in Table 1. We see that the lengths for the linear model are smaller than the ones in the logarithmic model. This is in agreement with the results obtained with stochastic gradient descent and shown in [1]. Finally, we summarize the cost values after parameter calibration in Table 2. The least values of every column are highlighted

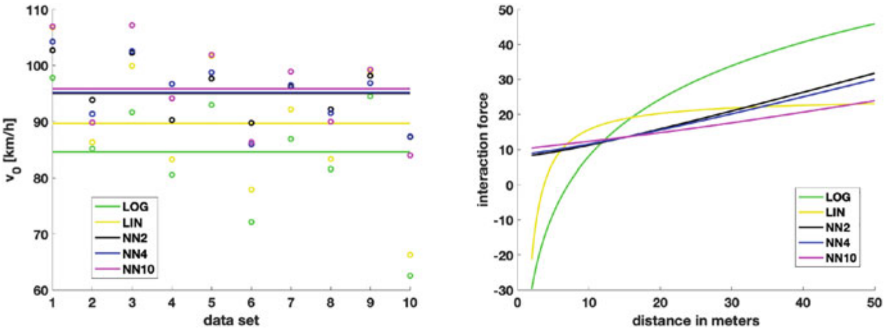


Fig. 1 Average velocities and forces resulting from the parameter calibration and learning process

[1] https://github.com/ctotzeck/NN-interaction.

Table 1 Car lengths (in m) estimated with the algorithm for the 10 data sets with the LWR-model with linear and logarithmic velocity

	1	2	3	4	5	6	7	8	9	10	average
Lin	3.5969	3.76	4.17	2.19	3.02	2.81	5.92	5.86	2.14	3.65	3.71
Log	7.15	7.21	8.05	8.17	6.19	5.00	8.10	8.46	5.63	6.91	7.09

Table 2 Values of the cost functional estimated with the algorithm for the 10 data sets with the LWR-model with linear and logarithmic velocity and the three different neural network approaches

	1	2	3	4	5	6	7	8	9	10	average
NN2	47.95	46.49	98.07	44.97	23.69	29.72	40.69	55.75	11.50	68.91	46.77
NN4	47.82	46.09	97.01	51.84	23.33	26.71	41.60	**55.29**	11.16	67.60	46.84
NN10	47.90	45.78	99.20	42.50	22.16	**24.40**	41.18	56.68	10.01	66.01	45.58
Lin	**44.41**	**41.29**	**93.73**	**30.86**	**19.00**	37.98	**38.00**	56.40	**8.18**	**46.24**	**41.61**
Log	53.53	50.31	109.36	65.24	26.50	52.93	38.09	58.22	14.54	52.75	52.15

in bold. It is obvious that the LWR model with linear force outperforms the other models. The results of the NN approaches are better than the ones of the LWR model with logarithmic force.

4.2.1 Comparison to Calibration with Stochastic Gradient Descent

In comparison to the parameter calibration based on the stochastic gradient descent method reported in [1], we find that the CBO approach finds better parameters for both LWR models. In fact, the resulting cost values are significantly smaller after the calibration with CBO. For the NN approaches the results are in good agreement. A clear decision in favour of the LWR approach or the NN ansatz was not possible based on the results of [1]. After the training with CBO the LWR with linear force seems to outperform all other approaches. We used NN with very simple structure here, it may be worth to test more sophisticated network structures in future work.

References

1. S. Göttlich and C. Totzeck, *Parameter calibration with stochastic gradient descent for interacting particle systems driven by neural networks*, Mathematics of Control, Signals, and Systems, online first, 2021.
2. C. Totzeck, *Trends in consensus-based optimization*, Active Particles, Volume 3, Springer International Publishing, 2022.
3. E. Kallo, A. Fazekas, S. Lamberty, and M. Oeser, *Microscopic traffic data obtained from videos recorded on a German motorway*, Mendeley Data, 2019. https://doi.org/10.17632/tzckcsrpn6.1
4. R. Pinnau, C. Totzeck, O. Tse, and S. Martin, *A consensus-based model for global optimization and its mean-field limit*, Math. Mod. Meth. Appl. Sci. **27**(1) (2017), 183–204
5. J. A. Carrillo, S. Jin, L. Li, and Y. Zhu, *A consensus-based global optimization method for high dimensional machine learning problems*, ESAIM: COCV **27** (2021) S5

Cancer Fingerprints by Topological Data Analysis

Ana Carpio

Abstract Topological data analysis has arisen has a promising tool to extract information on the structure of a wide variety of datasets. We analyze here its potential in two types of cancer studies. First, we compare times series of images from simulations of metastatic invasion in epithelial tissues. Calculating bottleneck distances of persistent diagrams we can characterize and classify the advancing interfaces of cellular aggregates. Second, we compare mRNA expression values for genes involved in cell cycles extracted from pancreas cancer tissue. We discuss how persistence information from different distances can provide insight on patient/gene clusters.

1 Introduction

Clinical and experimental studies of illness generate large amounts of data of a different nature. Consider cancer, for instance. Laboratory analyses of gene expression lead to large files containing measurements for different genes [15], see Fig. 1a. Instead, experimental observations of normal and malignant cells [9] yield time series of images, see Fig. 1b. Being able to extract meaningful information from large biomedical datasets, regardless of their nature, is a challenge that requires the development of adequate mathematical and computational tools.

Topological data analysis (TDA) furnishes a framework that provides dimensionality reduction and robustness to noise [2] when studying data clouds, with a certain independence with respect to the metrics selected. Recent studies have pointed out the potential of TDA in biological applications [8, 13, 16]. Biomedical data can be often be seen as point clouds in a space of dimension D. Whereas for images D is the spatial dimension, for gene expression datasets D is the number of patients or genes in the study. We will see how to use TDA to extract information in

A. Carpio (✉)
Universidad Complutense de Madrid, Madrid, Spain
e-mail: ana_carpio@mat.ucm.es

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_4

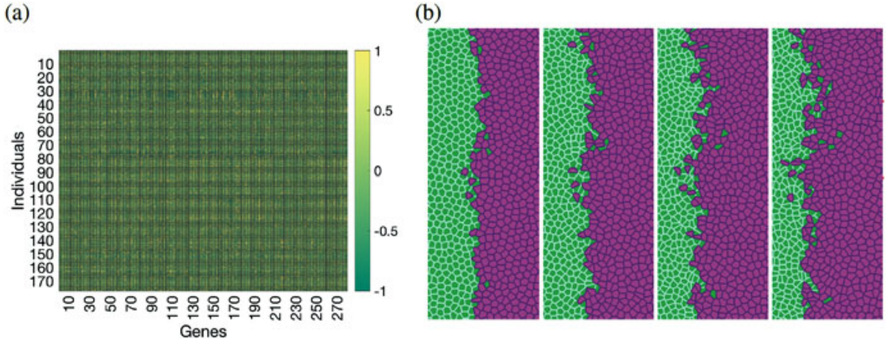


Fig. 1 (**a**) Heatmap showing normalized mRNA expressions for a collection of genes within a set of patients, data taken from [6]. (**b**) Snapshots from a numerical simulation showing the invasion of healthy (green) epithelial tissue by malignant (magenta) cells, reprinted from [1], see [9] for related experimental images

both settings. The paper is organized as follows. Section 2 applies TDA to classify automatically interfaces between healthy and malignant cells in two dimensional images. Section 3 proposes a topology based hierarchical clustering procedure for gene expression data. Finally, Sect. 4 summarizes our conclusions.

2 Classification of Interfaces

Competition between two different media (fluids, for instance) or populations is an ubiquitous phenomenon in many fields. Usually, an interface separating the two components forms. Being able to automatically characterize such interface is important to identify patterns or stages in biological applications. Given several images representing the evolution of fragmented interfaces, our strategy proceeds in the following steps [1]:

1. Extract from each image a point cloud X defining the interface.
2. Build a Vietoris-Rips filtration $V(X, r)$ for each point cloud based on the Euclidean distance, that is, a family of simplicial complexes formed joining by edges and triangles the points at a distance smaller than a variable parameter r, see [17].
3. Calculate the Betti numbers associated to each filtration: $betti_0(r)$ (number of components) and $betti_1(r)$ (number of holes) as the filtration parameter r varies.
4. For each identified component in each filtration, calculate the persistence intervals $[r_b, r_d]$, that is, the filtration parameter values at which it appears r_b (birth) and disappears r_d (death). They define the H_0 homology.
5. For each identified hole in each filtration, calculate the persistence intervals $[r_b, r_d]$. They define the H_1 homology.

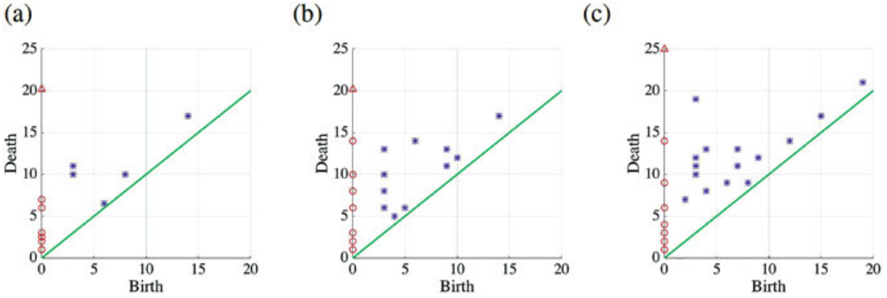


Fig. 2 Persistence diagrams representative of the initial, intermediate and late stages in the invasion process

6. Plot the persistence diagrams formed by the points (r_b, r_d) defining the persistence intervals for components and holes in each filtration, see Fig. 2.
7. Calculate the Bottleneck distance [11] between the H_1 persistence diagrams.
8. Use k-means or a hierarchical clustering [10] approach to group the interfaces in clusters according to the level of detail required.

For the simulation considered in Fig. 1b, a set of 12 images is classified by K-means in 3 groups: the first three frames correspond to initial stages in which the interface is close to an unfragmented smooth curve, the last two frames correspond to late stages of the invasion period with many fragments and interpenetration, while the remaining frames correspond to an intermediate stage in which fragments may detach and reattach, see Fig. 2.

The study of images involves point clouds in two or three dimensional spaces. Medical records containing the values of several variables monitorized over a collection of patients belong to higher dimensional spaces. Their study presents new difficulties.

3 Grouping Data

Gene studies in cancer patients have provided large amounts of information which may help to identify genetic features of sickness [15]. We consider here measurements of mRNA gene expression data for pancreas cancer available in [6], taken from the TCGA (the Cancer Genome Atlas) study. In this case, data take the form of numeric matrices $M = (m_{j,i})$ containing values for a collection of genes $i = 1, \dots, N$, from tissue samples corresponding to different patients $j = 1, \dots, J$.

The first step consists in normalizing the data. To do so [7], we calculate the means μ_i and standard deviations σ_i for each gene over the patients and compute the normalized values $\tilde{m}_{j,i} = \frac{m_{j,i}-\mu_i}{3\sigma_i}$. Then, we select a distance and a clustering strategy to group either patients using information from genes, or genes using information from patients.

3.1 Distance Selection

To compare genes or patients, we can use a number of distances [5]:

- The *Euclidean distance* between two columns or rows m^1 and m^2 is their distance as vectors in a D dimensional space $d(m^1, m^2) = \|m^1 - m^2\|_2$.
- The *Earth Mover's distance* (EMD) provides the minimum cost of turning one column (resp. row) into the other [13]

$$emd(m^1, m^2) = \frac{\sum_{k=1}^{D} \sum_{\ell=1}^{D} c_{k,\ell} d_{k,\ell}}{\sum_{k=1}^{D} \sum_{\ell=1}^{D} d_{k,\ell}},$$

  where $d_{k,\ell} = |m^1_k - m^2_\ell|$ is the ground distance, and $c_{k,\ell}$ minimizes the cost $\sum_{k=1}^{D} \sum_{\ell=1}^{D} c_{k,\ell} d_{k,\ell}$ subject to the constraints $c_{k,\ell} \geq 0,\ 1 \leq k, \ell \leq D$, $\sum_{k=1}^{D} \sum_{\ell=1}^{D} c_{k,\ell} = D$, $\sum_{k=1}^{D} c_{k,\ell} \leq 1,\ 1 \leq \ell \leq D$, $\sum_{\ell=1}^{D} c_{k,\ell} \leq 1,\ 1 \leq k \leq D$. The EMD identifies patterns regardless of their location. The distance between two patient profiles that are equal except for a peak about different genes would be small, which is inadequate as different genes may define different illnesses.
- Considering a set S of columns (resp. rows) m^1, m^2, …, m^L, the *Fermat* α-distance between any two of them relative to that set is [14]

$$d_{S,\alpha}(m^1, m^2) = \min\left\{ \sum_{\ell=1}^{k-1} \|y^{\ell+1} - y^{\ell}\|_2^{\alpha} \,\middle|\, (y_1, \dots, y_k) \text{ path from } m^1 \text{ to } m^2 \text{ in S} \right\},$$

  for any $\alpha > 1$. When $\alpha = 1$, we recover the Euclidean distance. The Fermat distance compares items in a set weighting information from all the other items in the same set, which is interesting when we want to compare gene profiles weighting information from cohorts of patients [3].

3.2 Distance and Topology Based Clustering

Figure 3 represents gene-gene and patient-patient distances for different gene (resp. patient) orderings. Regardless of the ordering, we can use such distance matrices in hierarchical clustering algorithms [10] and select a natural number of clusters based on inconsistency criteria [12]. Grouping genes (resp. patients) by their clusters we obtain the panels in Fig. 3, which uncover hidden relations in the data.

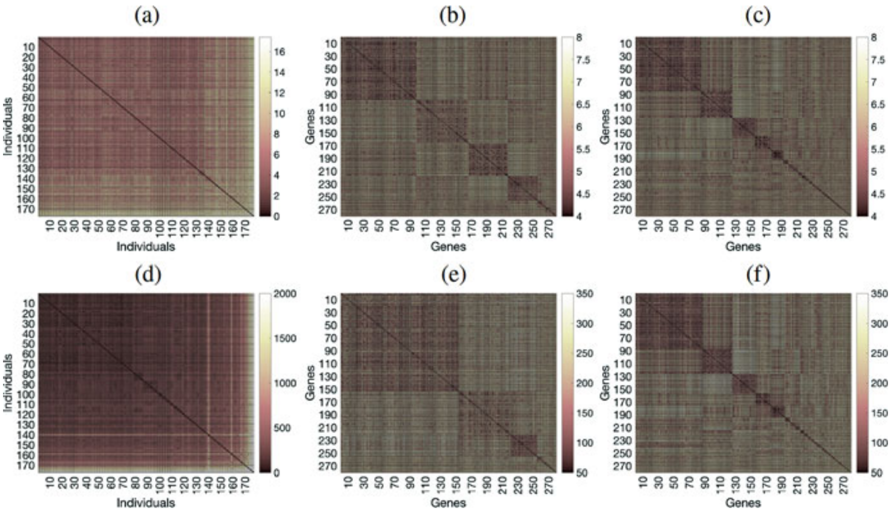


Fig. 3 Heatmaps representing the distance matrices for the set of data considered in Fig. 1a ordering patients (resp. genes) by cluster groups, as determined by hierarchical clustering with different distances: (**a**–**c**) Euclidean distances, (**d**–**f**) Fermat distances with $\alpha = 3$. (**a**) and (**d**) compare patients, while the rest compare genes. Panels (**a**–**b**), (**d**–**e**) use the natural number of clusters, as given by inconsistency studies. Instead, (**c**) and (**f**) use 36–37 clusters

Moreover, using any of these distances on the point cloud of patients $m_{j,\cdot} = (m_{j,1}, \dots, m_{j,N})$, $j = 1, \dots, J$, or the point cloud of patients $m_{\cdot,i} = (m_{1,i}, \dots, m_{N,i})$, $i = 1, \dots, N$, we can implement a similar procedure to that described in Sect. 2, only the distance changes. We construct a filtration, calculate the Betti numbers, as well as the persistence diagrams and intervals. With this information, we can compare datasets from different cancer types or patient studies to identify distinctive features and profiles. Moreover, the H_0 homology provides an additional clustering strategy, different from usual hierarchical clustering. For a fixed filtration parameter value, each component of the simplex constructed for that filtration value defines a cluster. As the filtration parameter varies, we have a topology based hierarchical clustering strategy. Figure 4 displays the same data as Fig. 1a when genes and patients are rearranged following the components of filtrations for a fixed filtration value.

4 Conclusions

We have discussed the potential of persistence studies based on different distances combined with clustering strategies to extract information from point clouds of data of medical interest. Applied to time series of images of cellular arrangements, it provides a tool to automatically classify specific image features. Applied to gene

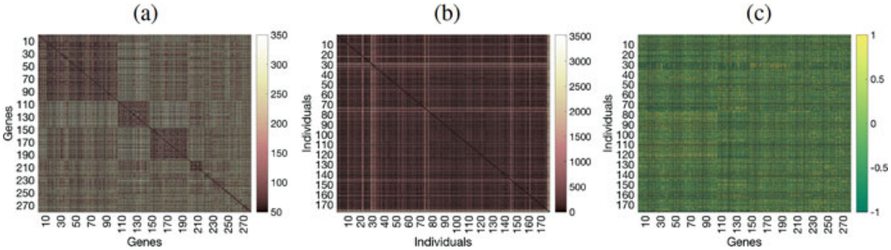


Fig. 4 Fermat distance reordered following H_0 clusters (**a**) for genes and (**b**) for patients. Panel (**c**) shows the data rearranged following the H_0 clusters

expression data, it opens new perspectives to gain a better understanding of hidden relations. Similar techniques could be exploited to study clinical data from other illnesses, immune disorders for instance [4].

Acknowledgments Research supported by Spanish MICINN grants MTM 2017-84446-C2-1-R and PID2020-112796RB-C21.

References

1. L.L. Bonilla, A. Carpio, C. Trenado, Tracking collective cell motion by topological data analysis, PLoS Comput Biol 16 (2020) e1008407.
2. G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2009) 255–308.
3. A. Carpio, L.L. Bonilla, J.C. Mathews, A.R. Tannenbaum, Fingerprints of cancer by persistent homology, bioRxiv 777169, 2019
4. A. Carpio, A. Simón, L.F. Villa, Clustering methods and Bayesian inference for the analysis of the time evolution of immune disorders, arXiv:2009.11531 2020
5. A. Carpio, A. Simón, A. Torres, L.F. Villa, Pattern recognition in data as a diagnosis tool, Journal of Mathematics in Industry 12 (2022) 3.
6. E. Cerami, J. Gao, U. Dogrusoz et al, The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data, Cancer Discov 2 (2012) 401–404.
7. Y. Chen, F. D. Cruz, R. Sandhu, A. L. Kung, P. Mundi, et al, Poediatric sarcoma data forms a unique cluster measured via the Earth Mover's Distance, Sci. Rep. 7 (2017) 7035.
8. M.R. McGuirl, A. Volkening, B. Sandstede, Topological data analysis of zebrafish patterns. Proc. Nat. Acad. Sci. 117 (2020) 5113–5124.
9. S. Moitrier, C. Blanch, S. Garcia, K. Sliogeryte et al., Collective stresses drive competition between monolayers of normal and Ras-transformed cells, Soft Matter 15 (2019) 537–545.
10. L. Kaufman, P.J. Rousseeuw, Finding groups in data: An introduction to cluster analysis, Hoboken: Wiley-Interscience, 1990.
11. M. Kerber, D. Morozov, A. Nigmetov, Geometry helps to compare persistence diagrams, ACM J. Exp. Algorithmics, 22 (2017) 1.4.
12. T. Kovacheva, A hierarchical clustering approach to find groups of objects, Proceedings of the IV Congress of Mathematicians, Macedonia; 2008. pp 359–373.
13. A.H. Rizvi, P.G. Camara, E.K. Kandror, T.J. Roberts et al., Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development, Nat. Biotech. 35 (2017) 551–560.

14. F. Sapienza, P. Groisman, M. Jonckheere, Weighted Geodesic Distance Following Fermat's Principle. Proc 6th International Conference on Learning Representations (ICLR), 2018.
15. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes, Nature 578 (2020) 82–93.
16. C. Topaz, L. Ziegelmeier, T. Halverson, Topological data analysis of biological aggregation models, PLoS ONE 10 (2015) e0126383.
17. A. Zomorodian, G. Carlsson, Computing persistent homology. Discrete and Computational Geometry, 33 (2002) 249–274.

Projected AQIF Parallel Algorithm for Solving EHL Line and Point Contact Problems: Parallel Computing

Peeyush Singh and Pravir K. Dutt

Abstract A novel parallel approach is developed for solving EHL line and point contact problems. The main motivation of algorithm comes from solving a discrete variational inequality problems on parallel computer by introducing a novel solver named as projected alternate quadrant interlocking factorization (PAQIF). The PAQIF has the property that when complementarity system

$$\begin{aligned} L_0 x &\geq b, \\ x &\geq 0, \\ x(L_0 x - b) &= 0 \end{aligned}$$

is banded with semibandwidth β_v, the space generated by $e_i., e_{n-i}; 1 \leq i \leq \beta_v$ is invariant under the transformation W^{-1}. Hence PAQIF is combined with partitioned scheme that renders a divide and conquer algorithm for solution of the banded linear complementarity system. The idea is extended to EHL problems by developing suitable preconditioner in the form of banded matrix.

1 Introduction

In a wide range of lubricated industrial devices studied, due to varying partial differential equations (PDEs) behaviour in Reynold's equation in the model (known as Elasto-hydrodynamic lubrication (EHL) see for examples [1, 2, 4]), depicting the pressure distribution and film thickness gap having considerable amount of difficulty

P. Singh (✉)
Department of Mathematics, VIT-AP University, Amaravati, India
e-mail: peeyush.singh@vitap.ac.in

P. K. Dutt
Department of Mathematics, IIT Kanpur, Kanpur, India
e-mail: pravir@iitk.ac.in

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_5

when the numerical simulation is done on serial computer. A very fine mesh is essential to capture inherited physics behind the model which generates a large memory requirement and computational complexity during the computation. Such a challenge can be compromised if discretized Reynold's equation is approximated in the form of a banded linear system during fix point iteration. Such banded linear systems often give rise to very large narrow banded linear systems which can be dense or sparse within the band. As result it is essential to develop robust parallel algorithms to meet the memory requirement and reduce the computational complexity by sharing the load on parallel computers. We discuss a novel parallel approach known as projected alternate quadrant interlocking factorization (PAQIF) to tackle the above mentioned extremities.

2 The Mathematical Model Problem

The mathematical formulation of the EHL problem consists of the set of nonlinear PDEs in the form of inequalities (see [1, 2] and [4] for more details) described as

$$\begin{aligned} \frac{\partial(\rho H)}{\partial x} - \frac{\partial}{\partial x}\Big(\epsilon \frac{\partial p}{\partial x}\Big) - \frac{\partial}{\partial y}\Big(\epsilon \frac{\partial p}{\partial y}\Big) &\geq 0 \quad \forall x, y \in \Omega \\ p(x, y) &\geq 0 \quad \forall x, y \in \Omega, \\ p(x, y)\Big[\frac{\partial(\rho H)}{\partial x} - \frac{\partial}{\partial x}\Big(\epsilon \frac{\partial p}{\partial x}\Big) - \frac{\partial}{\partial y}\Big(\epsilon \frac{\partial p}{\partial y}\Big)\Big] &= 0 \quad \forall x, y \in \Omega, \\ p(x, y) &= 0 \quad \forall x, y \in \partial\Omega. \end{aligned} \tag{1}$$

The elastic regime of the film thickness gap H between two contacting surfaces is governed by

$$H(p) = H_0 + \frac{x^2 + y^2}{2} + \frac{2}{\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{p(x^{'}, y^{'}) dx^{'} dy^{'}}{\sqrt{(x - x^{'})^2 + (y - y^{'})^2}}. \tag{2}$$

The dimensionless force balance equation are defined as follows

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x', y') dx' dy' = \frac{3\pi}{2}, \tag{3}$$

Here term ϵ is defined as

$$\epsilon = \frac{\rho H^3}{\eta \lambda},$$

where dimensionless viscosity η is defined as dimensionless density ρ and speed parameter λ

2.1 The PAQIF Algorithm

We consider the linear complementarity problem (LCP) define as below

$$\begin{aligned} LU(x) \geq f(x), &\quad x \in \Omega, \\ U(x) \geq 0, &\quad x \in \Omega, \\ U(x)^T.[LU(x) - f(x)] = 0, &\quad x \in \Omega, \\ U(x) = g(x), &\quad x \in \partial\Omega. \end{aligned} \tag{4}$$

Now we subdivide the LCP into r blocks linear sub-complementarity problem (LSCP) each of size n along the main diagonal such that $N = nr$, where r is the number of processors available. From Eq. (4), LSCP is expressed as

$$\begin{aligned} L_-^{(m)}U^{(m-1)}(x) + L_0^{(m)}U^{(m)}(x) + L_+^{(m)}U^{(m+1)}(x) \geq f^{(m)}(x), \quad m = 1, 2, \ldots, r \\ U^{(m)}(x) \geq 0 \\ U^{(m)}(x)^T.(L_-^{(m)}U^{(m-1)}(x) + L_0^{(m)}U^{(m)}(x) + L_+^{(m)}U^{(m+1)}(x) - f^{(m)}(x)) = 0, \end{aligned} \tag{5}$$

For each partition r, Eq. (5) can be reformulated as

$$\begin{aligned} L_0^{(m)}U^{(m)}(x) \geq f^{(m)}(x) - \begin{bmatrix} L_-^{(m)}U_L^{(m-1)}(x) \\ 0 \\ . \\ . \\ 0 \\ L_+^{(m)}U_F^{(m+1)}(x) \end{bmatrix}_{n\times 1} := f^{*(m)}(x), \quad m = 1, \ldots, r \\ U^{(m)}(x) \geq 0 \\ U^{(m)}(x)^T.(L_0^{(m)}U^{(m)}(x) - f^{*(m)}(x)) = 0, \end{aligned} \tag{6}$$

where $U_L^{(m-1)}(x)$ and $U_F^{(m+1)}(x)$ are $\beta_v \times 1$ vectors picked up from the last and first β_v components of the solution vector $U^{(m-1)}(x)$ and $U^{(m+1)}(x)$, respectively. Now we decouple the LSCP in Eq. (6) for parallel processors. Note that in Eq. (6) $f^{*(m)}(x)$ differs from $f^{(m)}(x)$ only in its first β_v and last β_v components. In order to factorize $L_0^{(m)}$ into $W_0^{(m)} Z_0^{(m)}$, we consider the space generated by

$$\mathrm{Span}_{1 \le i \le \beta_v} \{e_i, e_{n-i+1}\}$$

is invariant under the matrix $W_0^{(m)}$ (and so for $W^{(m)}{}_0^{-1}$), where

$$e_j := (0, 0, \dots, 0, 1_{j^{th} term}, 0, \dots, 0).$$

Let $[L_0^{(m)}]_{n \times n}$ (say $n = 2s$), $[W_0]_{n \times n}$ and $[Z_0]_{n \times n}$ matrices such that $L_0 = W_0 Z_0$, The above factorization can be proved that the method is stable for nonsingular diagonally dominant. Over all method is now outlined in brief as follows (see [3, 4] in details):

Step 1: For $m = 1, 2, \dots, r$ factorize in parallel

$$L_0^{(m)} = W_0^{(m)} Z_0^{(m)}$$

Step 2: For $m = 1, 2, \dots, r$ compute $Y^{(m)}$ in parallel

$$W_0^{(m)} Y^{(m)} = F^{(m)}$$

Step 3: For $m = 1, 2, \dots, r$ get inverse of $2\beta_v \times 2\beta_v$ matrix obtained by collecting first β_v and last β_v rows and columns of $W_0^{(m)}$ in parallel.

Step 4: Solve the reduced system from the subsystem by collecting first β_v and last β_v equations from each block. Then form normal equations, Solve system for $U_F^{(m)}$ and $U_L^{(m)}$, $m = 1, 2, \dots, r$.

Step 5: Project $U_F^{(m)}$ and $U_L^{(m)}$, $m = 1, 2, \dots, r$ into convex set K, where

$$K = \{p \in U : p \ge 0\}.$$

Step 6: For $m = 1, 2, \dots, r$ solve $U_M^{(m)}$ in parallel.

Step 7: Project $U_M^{(m)}$, $m = 1, 2, \dots, r$ into convex set K.

3 Numerical Results

We discretize the EHL model problem defined in Eqn (1) using finite difference method (see for example [4]). The domain decomposition method is used here for solving problem on parallel computers. We have used PAQIF algorithm during the fix point inner iteration process of the the computation. The speedup performance and efficiency plot of PAQIF algorithm is shown for varying grid points in Figs. 1 and 2 respectively. The converged pressure profile and gap plot are shown in Figs. 3

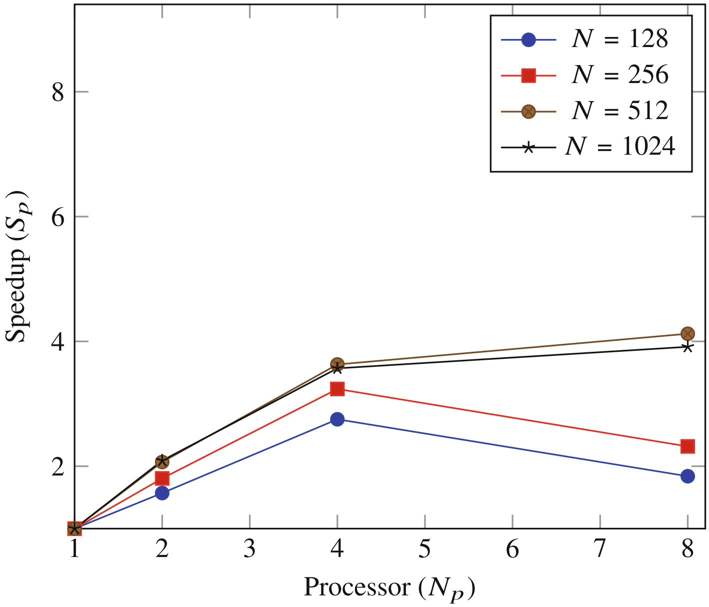


Fig. 1 Speedup plot for the cases $N = 128, 256, 512, 1024$, where bandwidth of matrix $\beta_v = 2$

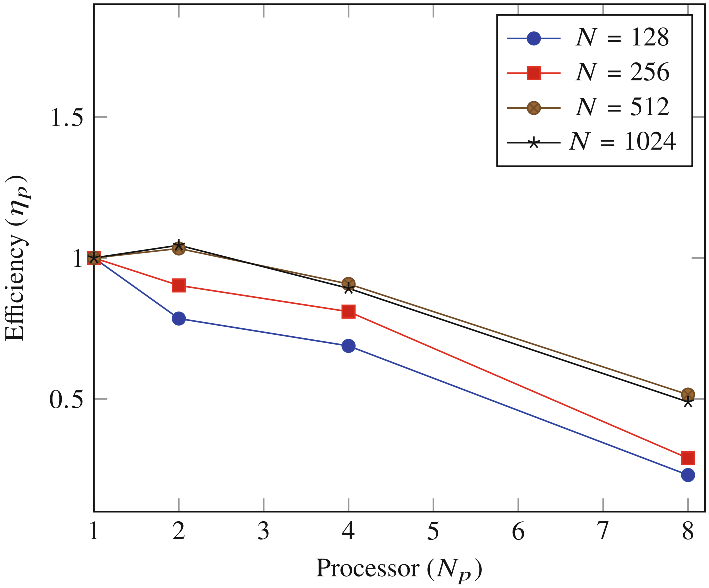


Fig. 2 Efficiency plot for the cases $N = 128, 256, 512, 1024$, where bandwidth of matrix $\beta_v = 2$

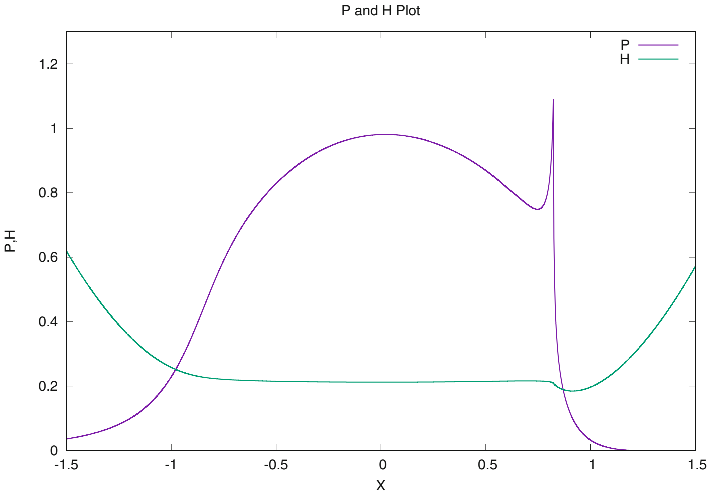


Fig. 3 EHL line contact, see [4]

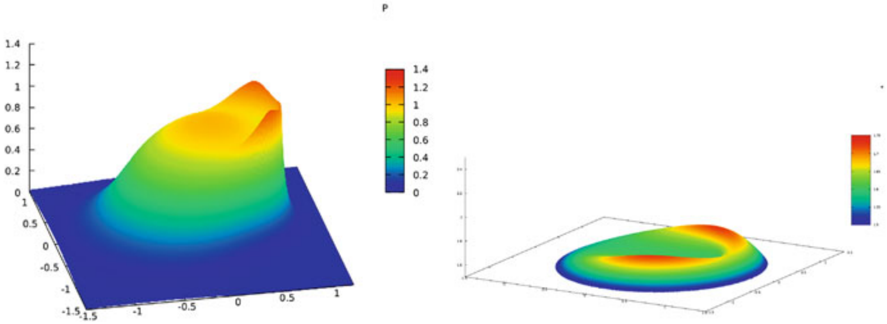


Fig. 4 Pressure P plot and 2-D Gap H plot for $M = 20$, $L = 10$, see [4]

and 4, respectively. We have performed all numerical computation on Dell Tower precision machine having processor specification Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz.

Acknowledgments First author got support from DST-SERB project no.PDF/2017/000202 and the Tata Institute of Fundamental Research, Centre for applicable mathematics,Bangalore, India.

References

1. P. Singh and P. Sinha, Interior-exterior penalty approach for solving elasto-hydrodynamic lubrication problem: Part I, Int. J. Numer. Anal. Model. 7(5), 695–731, 2020.
2. P. Singh and P. Sinha, Robust Numerical Solution for Solving Elastohydrodynamic Lubrication (EHL) Problems using Total Variation Diminishing (TVD) Approach, Commun. Math. Model. Appl. 4(2), 32–64, 2019.
3. S.C.S. Rao, P.K. Dutt, and M.K. Kadalbajoo, A Parallel Algorithm for Banded Linear System. Parallel Algorithm and Applications 14 (1999), 235–252.
4. P. Singh and P. Dutt, Total Variation Diminishing (TVD) method for Elastohydrodynamic Lubrication (EHL) problem on Parallel Computers. arxiv preprint 2008.03276, 2020.

Effectivity Analysis of Operator Splitting and the Average Method

Lívia Boda and István Faragó

Abstract In mathematics there are several problems which can be described by differential equations of a certain very complicated structure. Most of the time, we cannot produce the exact (analytical) solution of these problems, so we have to approximate them numerically by using some approximating method. In this paper we analyse one of such approximation methods, namely, operator splitting, which is a widely and successfully used method in numerical analysis. We introduce and demonstrate the method on a general Cauchy problem. In Sect. 1 of this paper we discuss the two most popular splitting methods, which are the sequential splitting (SS) and the Strang–Marchuk (SM) splitting, and describe the Average Method (AM) obtained by using splitting methods. Here we also discuss the possible reduction of the terms needed for the Average Method by using a matrix decomposition of pairwise commuting matrices.

In Sect. 2 we describe an aerodynamical model of flutter, which serves as our example problem. The advantage of the Average Method is shown in Sect. 3, where tables about runtimes and errors are given.

1 Operator Splitting and Average Method

We consider the following Cauchy-problem in $\mathbb{R}^m$:

$$\begin{cases} \dot{y}(t) = Ay(t) = \sum_{i=1}^{d} A_i y(t),\ t \in (0, T], \\ y(0) = y_0, \end{cases} \tag{1}$$

where $y\colon [0, T] \to \mathbb{R}^m$ is the unknown function, $y_0 \in \mathbb{R}^m$ is the given initial vector and $A_i \in \mathbb{R}^{m\times m}$ $(i = 1, \ldots, d)$ are given matrices.

L. Boda (✉) · I. Faragó
Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: boda.livia@ttk.bme.hu; istvan.farago@ttk.elte.hu

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_6

The exact solution of the Cauchy-problem (1) can be written directly as $y(t) = \exp(tA)y(0)$. Since the exact representation of the exponential matrix $\exp(tA)$ is typically impossible (or, at least, a very time-consuming task), our aim is to approximate the exact solution numerically by some suitable approximation of this exponential matrix on the grid

$$\omega_h = \left\{t_n = n \cdot h, h = \frac{T}{N}, \, n = 0, 1, \ldots, N\right\}. \tag{2}$$

We can do it by the so-called operator splitting, which means the following. We decompose the original (complex) problem into a series of simpler Cauchy problems, linked through their initial conditions. By applying this method it can be easier to find a numerical solution to the original problem.

The two most popular splitting methods include the sequential splitting (SS) and the Strang-Marchuk (SM) splitting. The algorithm of sequential splitting in case of two subproblems is as follows. In this case the decomposition of A is $A = A_1 + A_2$. If we use the sequential splitting to solve (1) on the grid ω_h, it means that the following two subproblems are solved in every step:

$$\begin{cases} \dot{y_1}(t) = A_1 y_1(t), \; t \in (t_i, t_{i+1}], \\ y_1(t_i) = x_{sp}(t_i), \end{cases} \tag{3}$$

$$\begin{cases} \dot{y_2}(t) = A_2 y_2(t), \;\; t \in (t_i, t_{i+1}], \\ y_2(t_i) = y_1(t_{i+1}). \end{cases} \tag{4}$$

where $i = 0, \ldots, N-1$, $x_{sp}(t_{i+1}) = y_2(t_{i+1})$ and $x_{sp}(t_0) = y_0$.

The main difference between the sequential and Strang-Marchuk splitting is that the latter computes values in the midpoints of the subintervals. The algorithm of SM splitting means solving the following subproblems:

$$\begin{cases} \dot{y}_1(t) = A_1 y_1(t), \; t \in \left(t_i, t_{i+\frac{1}{2}}\right], \\ y_1\left(t_i\right) = x_{sp}(t_i), \end{cases} \tag{5}$$

$$\begin{cases} \dot{y}_2(t) = A_2 y_2(t), \quad t \in \left(t_i, t_{i+1}\right], \\ y_2\left(t_i\right) = y_1\left(t_{i+\frac{1}{2}}\right), \end{cases} \tag{6}$$

$$\begin{cases} \dot{y}_1(t) = A_1 y_1(t), \qquad t \in \left(t_{i+\frac{1}{2}}, t_{i+1}\right], \\ y_1\left(t_{i+\frac{1}{2}}\right) = y_2\left(t_{i+1}\right). \end{cases} \tag{7}$$

where $i = 0, \ldots, N-1$, $x_{sp}(t_{i+1}) = y_1(t_{i+1})$ and $x_{sp}(t_0) = y_0$.

Remark 1 The sequential splitting is a first-order of consistency method, the Strang-Marchuk splitting is a second-order of consistency method.

As an alternative to the classical splitting methods, we introduce the Average Method with sequential splitting (AM_{SS}) which is based on the idea to divide the

Cauchy problem (1) into d subproblems, using sequential splitting in all possible ordering sequences. Then we define the numerical solution of each split subproblem, taking their arithmetic mean, and we define the new numerical solution in ω_h.

Let $\mathcal{P}^d$ denote the set of the permutations of the indices $\{1, 2, \dots, d\}$ and for $p = \{p_1, p_2, \dots, p_d\} \in \mathcal{P}^d$ we introduce the notation

$$\exp\{p_1, p_2, \dots, p_d\} = \exp(hA_{p_1})\exp(hA_{p_2}) \cdot \ldots \cdot \exp(hA_{p_d}). \tag{8}$$

Theorem 1 *Solving the Cauchy-problem* (1) *using sequential splitting for all possible permutations and then averaging the resulting numerical solutions yields a second-order method, i.e.*

$$\exp\big(h(A_1 + \ldots + A_d)\big) = \frac{1}{d!}\sum_{p\in\mathcal{P}^d} \exp\{p_1, p_2, \dots, p_d\} + O(h^3). \tag{9}$$

So instead of using a second-order method once, we just use a first-order method more than once and we get a second-order numerical solution. Hence the main advantage is that a first-order method requires less computational demand than a second-order numerical method. However using the AM_{SS} method to solve Cauchy problem (1), we have to calculate $d!$ numerical solutions. Even with a relatively small value of d, we have to produce many numerical solutions and the computational demand may increase greatly.

If we find a decomposition for Cauchy problem (1) that includes commuting matrices, the number of subproblems can be significantly reduced. Let $A = A_1 + A_2 + \ldots + A_d$, and suppose that $\exists i, j \in \mathbb{N}$, and $i \neq j$ such that $[A_i, A_j] = 0$. Then instead of all the $d!$ permutations, we have $d! - (d-1)! = (d-1)(d-1)!$ elements. If the decomposition includes more commuting pairs of matrices, the reduction might be more significant. The other advantage of the Average Method is that the $d!$ numerical solutions can be independently calculated, i.e. the computation is parallelizable.

2 Application to the Aerodynamics

We investigated the efficiency of the Average Method on a physical problem which describes the aerodynamics of an airplane wing. The model is based on a wind tunnel experiment in which the lift force of an airplane wing was examined as a function of the inclination of the wing. This piecewise-linear model of flutter was investigated in [2]. Motivated by this model, we consider the 4-dimensional Cauchy problem

$$\begin{cases} \dot{\mathbf{x}}(t) &= A_k\mathbf{x}(t), \\ \mathbf{x}(0) &= \mathbf{x}_0. \end{cases} \tag{10}$$

Table 1 Parameters of the model

Parameter	c_0	c_1	c_2	p_1	p_2	p_3	p_4
Value	5.932	-6.846	2.662	0.1485	0.0147	0.0540	0.2748

where the affine model equations contain the three system matrices ($k = 0, 1, 2$)

$$A_k = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & -(p_1 + p_2\mu c_k) & -\mu^2 c_k p_2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & c_k\mu & -(p_4 - c_k\mu^2) & -p_3 \end{pmatrix},$$

with the model parameters given in Table 1, and $\mu \in (0, \infty)$ represents the nondimensional wind speed.

We analyzed several decompositions of matrix A_k, the most important of them being the following, which contains commuting matrices:

$$A_k = A_{k_{(1)}} + A_{k_{(2)}} + A_{k_{(3)}}, \tag{11}$$

where

$$A_{k_{(1)}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & -(p_1 + p_2\mu c_k) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_{k_{(2)}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -(p_4 - c_k\mu^2) & -p_3 \end{pmatrix},$$

$$A_{k_{(3)}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\mu^2 c_k p_2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & c_k\mu & 0 & 0 \end{pmatrix}.$$

Clearly, that matrices $A_{k_{(1)}}$ and $A_{k_{(2)}}$ are commuting matrices, therefore:

$$\exp\left(hA_{k_{(1)}}\right) \cdot \exp\left(hA_{k_{(2)}}\right) = \exp\bigl(h(A_{k_{(1)}} + A_{k_{(2)}})\bigr). \tag{12}$$

Then we introduce the notation

$$A_{k_{(4)}} = A_{k_{(1)}} + A_{k_{(2)}} = A_{k_{(2)}} + A_{k_{(1)}}. \tag{13}$$

Solving the Cauchy problem (1) using the AM_{SS} method with decomposition (11), which includes commuting matrices, and using property (12) and notation (13), the number of subproblems can be reduced from six to four, and we have to calculate the following four numerical solutions:

$$x_1(h) = \exp(hA_{k_{(1)}})\exp(hA_{k_{(3)}})\exp(hA_{k_{(2)}}) \cdot x_0, \tag{14}$$

$$x_2(h) = \exp(hA_{k_{(2)}})\exp(hA_{k_{(3)}})\exp(hA_{k_{(1)}}) \cdot x_0, \tag{15}$$

$$x_3(h) = \exp(hA_{k_{(4)}})\exp(hA_{k_{(3)}}) \cdot x_0, \tag{16}$$

$$x_4(h) = \exp(hA_{k_{(3)}})\exp(hA_{k_{(4)}}) \cdot x_0, \tag{17}$$

where (14), (15) have three subproblems, and in case of (16), (17) the number of subproblems was reduced from three to two which further simplifies the solution process and reduces computational demand, too. Then based on Theorem 1, we have

$$\exp\left(h(A_{k_{(1)}} + A_{k_{(2)}} + A_{k_{(3)}})\right) = \frac{x_1(h) + x_2(h) + 2x_3(h) + 2x_4(h)}{6} + \mathcal{O}(h^3). \tag{18}$$

3 Numerical Results

During the numerical implementation, the numerical solutions (14)–(17) were computed using sequential splitting, which has first order, and the subproblems were solved using the first-order explicit Euler method. Then averaging these first-order solutions we get a second-order numerical solution. The essence of the AM$_{SS}$ method is that we have to implement some first-order approximating methods, which we can easily implement, then the average of first-order numerical solutions should be taken, which is not a very expensive operation, either, then we get a second-order method.

The numerical solutions (14)–(17) can be independently calculated, i.e. the computation is parallelizable. When we have four processors, we can compute the solutions (14)–(17) at the same time. We can simulate the parallel run as follows. Consider that we have four processors to calculate the numerical solutions in parallel. We can see the runtimes of every calculation of the solutions (14)–(17) in Table 2. We can calculate the whole runtime as follows: choose the maximum of the four runtimes (red coloured) and then add the runtime of the averaging. The last column of Table 2 shows the full runtime of the Average Method in case of four processors.

Now we consider the case where three processors are available to solve system (10) using the AM$_{SS}$ method. In this case the main problem is to partition the subproblems well. On the one hand we saw in Sect. 2 that in case of solutions (14) and (15) there are three subproblems with matrices $A_{k_{(1)}}$, $A_{k_{(2)}}$ and $A_{k_{(3)}}$ while in case of solutions (16) and (17) there are only two subproblems with matrices $A_{k_{(3)}}$ and $A_{k_{(4)}}$. And on the other hand Table 2 shows that solutions (16) and (17) can be computed faster than solutions (14) and (15). Therefore, it is reasonable to partition

Table 2 Runtimes (in seconds) of the AM$_{SS}$ during a parallel run with four processors

h	x_1	x_2	x_3	x_4	Average	Runtime
1	$7.20 \cdot 10^{-5}$	$6.12 \cdot 10^{-5}$	$2.10 \cdot 10^{-5}$	$2.22 \cdot 10^{-5}$	$1.02 \cdot 10^{-6}$	$7.30 \cdot 10^{-5}$
0.1	$8.24 \cdot 10^{-5}$	$8.90 \cdot 10^{-5}$	$7.56 \cdot 10^{-5}$	$7.14 \cdot 10^{-5}$	$2.21 \cdot 10^{-5}$	$1.11 \cdot 10^{-4}$
0.01	$8.18 \cdot 10^{-3}$	$7.52 \cdot 10^{-3}$	$1.71 \cdot 10^{-3}$	$2.03 \cdot 10^{-3}$	$5.62 \cdot 10^{-4}$	$8.74 \cdot 10^{-3}$
0.001	$4.02 \cdot 10^{-2}$	$3.89 \cdot 10^{-2}$	$1.37 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$	$9.72 \cdot 10^{-4}$	$4.11 \cdot 10^{-2}$

Table 3 Runtimes (in seconds) of AM$_{SS}$ during a parallel run with three processors

h	x_1	x_2	x_3 and x_4	Average	Runtime
1	$7.20 \cdot 10^{-5}$	$6.12 \cdot 10^{-5}$	$4.32 \cdot 10^{-5}$	$4.02 \cdot 10^{-6}$	$7.30 \cdot 10^{-5}$
0.1	$8.24 \cdot 10^{-5}$	$8.90 \cdot 10^{-5}$	$1.47 \cdot 10^{-4}$	$2.21 \cdot 10^{-5}$	$1.69 \cdot 10^{-4}$
0.01	$8.18 \cdot 10^{-3}$	$7.52 \cdot 10^{-3}$	$3.74 \cdot 10^{-3}$	$5.62 \cdot 10^{-4}$	$8.74 \cdot 10^{-3}$
0.001	$4.02 \cdot 10^{-2}$	$3.89 \cdot 10^{-2}$	$2.82 \cdot 10^{-2}$	$9.72 \cdot 10^{-4}$	$4.11 \cdot 10^{-2}$

Table 4 Runtimes (in seconds) of AM$_{SS}$ during a parallel run with two processors

h	x_1 and x_2	x_3 and x_4	Average	Runtime
1	$9.31 \cdot 10^{-5}$	$8.34 \cdot 10^{-5}$	$1.02 \cdot 10^{-6}$	$9.40 \cdot 10^{-5}$
0.1	$1.58 \cdot 10^{-4}$	$1.60 \cdot 10^{-4}$	$2.21 \cdot 10^{-5}$	$1.82 \cdot 10^{-4}$
0.01	$9.89 \cdot 10^{-3}$	$9.55 \cdot 10^{-3}$	$5.62 \cdot 10^{-4}$	$1.04 \cdot 10^{-2}$
0.001	$5.39 \cdot 10^{-2}$	$5.34 \cdot 10^{-2}$	$9.72 \cdot 10^{-4}$	$5.48 \cdot 10^{-2}$

as follows: solutions (14) and (15) are computed by two separate processors, and solutions (16) and (17) are computed one after the other by the third processor. Table 3 shows the runtimes of this case.

It is worth examining the case where we have two processors to compute the numerical solution of (10). Similarly to the three-processor case, proper partitioning will be the main task. The most reasonable partition is as follows: calculate solutions (14) and (16) one after the other with one processor, meanwhile solutions (15) and (17) can be calculated one after the other using the other processor. In this case Table 4 shows the runtimes.

And in order to see the practical usefulness of the AM$_{SS}$ method, we solved the whole Cauchy-problem (10) without any splitting process using the improved Euler method, which is the same second-order method as the AM$_{SS}$ method, and we compared the runtime of the AM$_{SS}$ with two, three and four processors with the runtime of the improved Euler method. Table 5 shows this comparison and we can see that on average, the AM$_{SS}$ method is one-two orders of magnitude faster than the improved Euler method.

Table 6 shows the comparison of errors in case of AM_{SS} and the improved Euler method. It can be seen the second-order convergence in both cases, furthermore the error is approximately the same in both cases.

Table 5 Comparison of runtimes (in seconds) in case of AM_{SS} with two, three and four processors and the improved Euler method

h	AM_{SS} + 2 proc.	AM_{SS} +3 proc.	AM_{SS} + 4 proc.	Improved Euler
1	$9.40 \cdot 10^{-5}$	$7.30 \cdot 10^{-5}$	$7.30 \cdot 10^{-5}$	$8.18 \cdot 10^{-3}$
0.1	$1.82 \cdot 10^{-4}$	$1.69 \cdot 10^{-4}$	$1.11 \cdot 10^{-4}$	$1.96 \cdot 10^{-2}$
0.01	$1.04 \cdot 10^{-2}$	$8.74 \cdot 10^{-3}$	$8.74 \cdot 10^{-3}$	$8.44 \cdot 10^{-2}$
0.001	$5.48 \cdot 10^{-2}$	$4.11 \cdot 10^{-2}$	$4.11 \cdot 10^{-2}$	$1.13 \cdot 10^{0}$

Table 6 Comparison of errors in case of AM_{SS} and the improved Euler method

h	AM_{SS} method	Improved Euler
1	$2.08 \cdot 10^{-2}$	$4.77 \cdot 10^{0}$
0.1	$2.44 \cdot 10^{-4}$	$2.04 \cdot 10^{-4}$
0.01	$2.44 \cdot 10^{-6}$	$2.01 \cdot 10^{-6}$
0.001	$2.44 \cdot 10^{-8}$	$2.01 \cdot 10^{-8}$

Acknowledgments This research has been supported by the National Research, Development and Innovation Office - NKFIH, grant no. K137699.

References

1. Faragó, I., Havasi, Á.: Operator splittings and their applications. In: Nova Science Publishers., (2009)
2. Kalmár-Nagy, T., Csikja, R., Elgohary, T.A.: Nonlinear analysis of a 2-dof piecewise linear aeroelastic system. Nonlinear Dynamics, vol. 85, no. 2, pp. 739–750 (2016)

Numerical Solutions of Boundary Value Problems Using the Carleman Linearisation Method

Gabriella Svantnerné Sebestyén

Abstract In this article we apply the Carleman linearisation method to solve boundary value problems. This method transforms sets of polynomial ordinary differential equations into an infinite dimensional linear system. We investigate the Carleman linearisation method to two-point boundary value problems and we have also analysed the error of this method through an example.

1 Introduction

In this article we investigate the numerical solution of two-point boundary value problems with the Carleman linearisation method. Different phenomena in physics or in engineering can be modelled by boundary value problems for example fluid dynamics and linear elasticity [1], [3]. In this article we consider the following two-point boundary value problem

$$\begin{aligned} x''(t) &= f(t, x(t), x'(t)), \quad t \in (a, b); \\ x(0) &= \alpha, \quad x(b) = \beta, \end{aligned} \tag{1}$$

where $x = x(t) : \mathbb{R} \to \mathbb{R}$ is the unknown function, $f : \mathbb{R}^3 \to \mathbb{R}$ in general is an nonlinear function and α and β are given numbers.

Usually we can not determine the exact solution of boundary value problems so we use numerical methods. Such methods are the shooting method and the finite difference method [2].

In the following we apply the Carleman linearisation method to solve boundary value problems. First we outline the method in general and after that we show the application to boundary value problems.

G. S. Sebestyén (✉)
Budapest University of Technology and Economics, Budapest, Hungary

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_7

2 Carleman Linearisation Method to Two-Point Boundary Value Problems

The Carleman linearisation is a method for solving nonlinear autonomous differential equations [4]. This approach is based on truncates an infinite-dimensional linear system and omits the higher-order terms. Let us consider the following system of differential equations with the power of the functions $x(t)$ and $y(t)$ in the following way

$$\begin{aligned} x'(t) &= \textstyle\sum_{k+l\geq 1}^{\infty} a_{k,l} x^k(t) y^l(t), \\ y'(t) &= \textstyle\sum_{k+l\geq 1}^{\infty} b_{k,l} x^k(t) y^l(t), \end{aligned} \tag{2}$$

with initial conditions

$$x(0) = x_0, \quad y(0) = y_0. \tag{3}$$

We introduce the functions

$$\mathbf{u}\,(t)^{[j]} = (x^j(t), x^{j-1}(t)y(t), \ldots, x^2(t)y^{j-2}(t), x(t)y^{j-1}(t), y^j(t))^{\mathrm{T}}, \tag{4}$$

$j = 1, \ldots, N$, containing the elements $x^{j-p}(t)y^p(t)$, $p = 0, 1, ..., j$. We introduce the vector

$$\mathbf{v}(t) = \left(\mathbf{u}^{[1]}(t),\ \mathbf{u}^{[2]}(t), \ldots, \mathbf{u}^{[N]}(t)\right)^{\mathrm{T}}, \tag{5}$$

then the Carleman embedded system has the form

$$\frac{d}{dt}\mathbf{v}\,(t) = \mathbf{C}_N \mathbf{v}(t) + \mathcal{O}\left(\mathbf{u}(t)^{[N+1]}\right), \tag{6}$$

where $\mathbf{C}_N$ is the Nth order Carleman matrix. By omitting the higher order terms $\mathcal{O}\left(\mathbf{u}^{[N+1]}(t)\right)$, then the Carleman linearized system is the following system of ordinary differential equations

$$\frac{d}{dt}\mathbf{v}(t) = \mathbf{C}_N \mathbf{v}(t). \tag{7}$$

known initial condition from (3), we have

$$\mathbf{v}_0 = \left(x_0,\ y_0,\ x_0^2,\ x_0 y_0,\ y_0^2, \ldots\right)^{\mathrm{T}} \tag{8}$$

and hence the exact solution of equation (7)–(8) has the form

$$\mathbf{v}(t) = e^{\mathbf{C}_N t}\mathbf{v}_0 \tag{9}$$

which will be considered as an approximation to the exact solution of the original problem (2)–(3). When we apply the previous method to two-point boundary value problems, at the first step we replace the second order differential equation to a first order system of differential equations in the following way

$$\begin{aligned} x'(t) &= y(t) \\ y'(t) &= f(t, x(t), y(t)). \end{aligned} \tag{10}$$

At the point zero we do not know the derivative value of the unknown function so we assume that $x'(0) = y(0) = c$, where $c \in \mathbb{R}$ is a given number. Then problem can be written as a first order problem with the following initial conditions

$$\begin{aligned} x'(t) &= y(t),\ x(0) = \alpha \\ y'(t) &= f(t, x(t), y(t)),\ y(0) = c. \end{aligned} \tag{11}$$

That means when we apply the Carleman linearisation method to the system (11) then the solution depends on the choice of the parameter c and it has the form

$$\mathbf{v}(t, c) = e^{\mathbf{C}_N t}\mathbf{v}_0(c). \tag{12}$$

At point b we know the value of the unknown function so we have to determine parameter c from the relationship

$$\mathbf{v}(b, c) = e^{\mathbf{C}_N b}\mathbf{v}_0(c) = \beta. \tag{13}$$

Hence, introducing the function

$$g(c) = e^{\mathbf{C}_N b}\mathbf{v}_0(c) - \beta \tag{14}$$

our aim is to solve the nonlinear equation $g(c) = 0$. Typically we use an approximation method e.g. the Newton-method to solve this nonlinear equation.

3 Numerical Simulations and Results

In this section we illustrate the Carleman linearisation method through an example and we analyse the error of the method.

3.1 Problem Settings

We consider the two-point boundary value problem:

$$\begin{aligned} x''(t) - 6x^2(t) &= 0 \\ x(0) &= 1 \\ x(0.5) &= 4. \end{aligned} \tag{15}$$

Problem (15) can be written to the following initial value problem

$$\begin{aligned} x'(t) = y, &\quad x(0) = 1 \\ y'(t) = 6x^2, &\quad y(0) = c \end{aligned} \tag{16}$$

where $c \in \mathbb{R}$ is some fixed (guessed) number. We apply the Carleman linearisation method to solve problem (16).

When $N = 1$, we solve the following first order system

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{C_1} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 6x^2 \end{pmatrix}. \tag{17}$$

The solution has the form $\mathbf{v}(t, c) = e^{C_1 t}\mathbf{v}_0(c)$, where $\mathbf{v}_0 = (1, c)^T$. Hence, the solution is

$$\mathbf{v}(t, c) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ c \end{pmatrix} = \begin{pmatrix} 1 + ct \\ c \end{pmatrix} \tag{18}$$

and therefore the solution of the problem is $x(t) = 1 + ct$. If we use the boundary value $x(0.5) = 4$,

$$c = 6. \tag{19}$$

When $N = 2$, we solve the following first order system

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \\ \frac{dx^2}{dt} \\ \frac{d(xy)}{dt} \\ \frac{dy^2}{dt} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{C_2} \begin{pmatrix} x \\ y \\ x^2 \\ xy \\ y^2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 6x^3 \\ 12x^2y \end{pmatrix}. \tag{20}$$

The solution has the form $\mathbf{v}(t,c) = e^{C_2 t}\mathbf{v}_0$, where $\mathbf{v}_0 = \left(1, c, 1, c, c^2\right)^T$. Let $A(t) := e^{C_2 t}$ be the exponential of the matrix C_2. We can determine the value of function x from the following relationship

$$x(t) = A_{11}(t) + A_{12}(t)c + A_{13}(t) + A_{14}(t)c + A_{15}(t)c^2. \tag{21}$$

From the boundary condition at the point we get

$$x(0.5, c) = 4 \tag{22}$$

which is a nonlinear equation for the parameter c. Solving the equation $g(c) = 0$ by using the Newton method, we obtain

$$c_{n+1} = c_n - \frac{A_{11}(0.5) + A_{12}(0.5)c_n + A_{13}(0.5) + A_{14}(0.5)c_n + A_{15}(0.5)c_n^2 - 4}{A_{12}(0.5) + A_{14}(0.5) + 2A_{15}(0.5)c_n} \tag{23}$$

iteration $n = 0, 1, \ldots$ where, $c_0 \in \mathbb{R}$ is a given number.

When N tends to infinity, we apply the Carleman linearisation method similarly. The solvable problem is larger and more complicated.

3.2 Numerical Results

In the following we examine the Carleman linearisation on the previous example. We use the following notations:

- c_0 is the initial value of the iteration,
- kmax is the number of the iteration,
- etol is the error tolerance,
- N is the Nth Carleman matrix.

We start the Newton method from value c_0 and use the Carleman iteration with different N values. In Table 1, we can see that value c tends to the exact solution ($c = 2$), when N increases.

Table 1 The approximation of value c.

c_0	kmax	etol	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
0	10	0.001	6	3	2.4444	2.3154	2.2855	2.2788	2.2773
1	10	0.001	6	2.8077	2.2500	2.1096	2.0766	2.0686	2.0668
1.5	10	0.001	6	2.75	2.2024	2.0600	2.0270	2.0188	2.0168
2	10	0.001	6	2.7143	2.1818	2.0424	2.0105	2.0025	2.0006
3	10	0.001	6	2.6969	2.1803	2.0423	2.0104	2.0025	2.0006

Table 2 Error in maximum norm

c_0	N = 1	N = 2	N = 3
0	0.7597	0.4023	0.2977
1	0.7597	0.3109	0.1270
1.5	0.7597	0.2850	0.0941
2	0.7597	0.2692	0.0810
3	0.7597	0.2616	0.0801

In Table 2, we can that the difference between the exact solution and the numerical solution in maximum norm.

We have seen that the Carleman linearisation method can be applied to solve two-point boundary value problems. We have seen through an example that when N tends to infinity then the error of then method tends to zero and the approximation of value c tends to the exact value of the derivative of the solution.

In the future we examine the application of the Carleman linearisation to partial differential equations and examine the error of the method.

References

1. Barber, Jr.: *Elasticity.* Kluwer Academic Publishers, Dordrecht, The Netherlands (2002)
2. Baxley, J. V., Nonlinear Two Point Boundary Value Problems in Ordinary and Partial Differential Equations (Everrit ,W.N. and Sleeman , B.D. Eds.). 46–54, Springer-Verlag, (1981)
3. Ebaid, A.: Approximate analytical solution of a nonlinear boundary value problem and its application in fluid mechanics, Zeitschrift fur Naturforschung A, vol. 66, no. 6–7, pp. 423–426, (2011)
4. Kowalski, Krzysztof, Steeb, Willi-Hans: Nonlinear Dynamical Systems and Carleman Linearization, World Scientific Publishing Co. Pte. Ltd., (1991)

Immersed Boundary Models of Biofilm Spread

Ana Carpio and Rafael González-Albaladejo

Abstract We propose an immerse boundary approach for the dynamics of active contours in flows. When the active contours represent bacterial boundaries, we couple this system with dynamic energy budget models of cell metabolism for the evolution of the cell boundaries, informed by reaction-diffusion systems for the relevant concentration fields. Numerical simulations illustrate the evolution of incipient biofilms formed by clusters of spherical bacteria in two dimensions.

1 Introduction

Immersed boundary (IB) methods [13, 14] provide efficient tools to handle fluid/structure interactions in many applications. Our goal here is to adapt them to describe the behavior of cellular systems such as bacterial biofilms, in which the structures are cell membranes. Biofilms are bacterial aggregates encased in a self-produced polymeric matrix which grow on moist surfaces [6] and are responsible for most hospital acquired infections [8]. Many models have been developed to study their behavior, focusing on different aspects: continuous models [17], agent based descriptions [7, 10, 11, 18, 20] and hybrid models combining both [2, 16]. Immersed boundary methods have already been used to study finger deformation [20], viscoelastic behavior [19] and attachment of bacteria [4] in flows. Active cellular contours have been addressed by removing the incompressibility constraint and including inner sources [12]. Applications to multicellular tissues

A. Carpio
Universidad Complutense de Madrid, Madrid, Spain
e-mail: ana_carpio@mat.ucm.es

R. González-Albaladejo (✉)
Universidad Complutense de Madrid, Madrid, Spain

Instituto Gregorio Millán, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: rafael09@ucm.es

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_8

[5, 15] consider closely packed deformable contours attached to each other [5, 15]. However, bacterial biofilms are formed by rigid shapes which remain at a certain distance. When biofilms grow in flows, we usually have scattering bacteria in large polymer fractions. Instead, we consider here incipient biofilms spreading on surfaces, in which the volume fraction of polymeric matrix keeping cells together is small [17]. We propose a computational model that combines an IB description of cellular arrangements and mechanical interactions with a dynamic energy budget (DEB) representation of bacterial activity and chemical processes. Simple tests on clusters of spherical bacteria illustrate its potential to investigate cell arrangements and interaction with flows.

2 Immerse Boundary Model for Active Boundaries

Immersed boundary models are usually formulated for 'inert' boundaries whose shape changes as a result of the interaction with the fluid, keeping a fixed size. Cells are 'active' boundaries, whose size and number changes. Let us explain how this affects the standard IB equations. Given a region Ω and a boundary Γ immersed in it, the fluid-structure interaction is described by the incompressible Navier-Stokes equations set in Ω [13, 14]:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = \nu \Delta \mathbf{u} - \frac{1}{\rho} \nabla p + \frac{1}{\rho} \mathbf{f} - \frac{\alpha}{\rho} \mathbf{u}, \quad \text{div}(\mathbf{u}) = 0, \tag{1}$$

where $\mathbf{u}(\mathbf{x}, t)$, $p(\mathbf{x}, t)$ and $\mathbf{f}(\mathbf{x}, t)$ are the fluid velocity, fluid pressure and external force density. The parameters ρ, $\nu = \mu\rho$ and α denote the fluid density, kinematic viscosity and friction coefficient, respectively. We enforce periodic boundary conditions for the fluid, which allows to use fast solvers based on fast Fourier transforms [13, 14], and place Γ far from the boundaries to allow for free growth while reducing boundary effects. The force $\mathbf{f}(\mathbf{x}, t)$ created by Γ on the fluid is

$$\mathbf{f}(\mathbf{x}, t) = \int_\Gamma \mathbf{F}(\mathbf{q}, t) \delta(\mathbf{x} - \mathbf{X}(\mathbf{q}, t)) \, d\mathbf{q}. \tag{2}$$

In practice, the delta function δ is replaced for computational purposes with adequate regularizations [13, 14]. $\mathbf{X}(\mathbf{q}, t)$ is the parametrization of Γ, and $\mathbf{F}(\mathbf{q}, t)$ is the force density on it. The integration parameters $\mathbf{q}$ represent angles. When several cells are present, we work with several parametrizations $\mathbf{X}_1, \dots, \mathbf{X}_N$.

The evolution equation for Γ follows correcting the no-slip condition

$$\frac{\partial \mathbf{X}}{\partial t} = \int_\Omega \mathbf{u}(\mathbf{x}, t) \delta(\mathbf{x} - \mathbf{X}(\mathbf{q}, t)) \, d\mathbf{x} + \lambda \big((\mathbf{F}_g \cdot \mathbf{n}) \mathbf{n} + \mathbf{F}_{ext} \big), \quad \lambda > 0, \tag{3}$$

with the contribution of growth $\mathbf{F}_g$ and external forces $\mathbf{F}_{ext}$. In practice, $\mathbf{F} = \mathbf{F}_e + \mathbf{F}_g + \mathbf{F}_{ext}$. Elastic forces $\mathbf{F}_e$ within the IB are tangent to the outer normal $\mathbf{F}_e \cdot \mathbf{n} = 0$. In two dimensions, and assuming the boundaries are formed by springs parametrized by the angle θ, $\mathbf{F}_e = \frac{\partial}{\partial \theta}\left(K \frac{\partial \mathbf{X}}{\partial \theta}\right)$, for an elastic constant K [13, 14]. Standard IB approaches set $\mathbf{F}_g = \mathbf{F}_{ext} = 0$ and $\alpha = 0$. Here, the friction parameter $\alpha > 0$ represents the effect of the polymeric matrix enveloping bacteria and hindering their motion. The growth forces are included since they modify the size of Γ. We set them proportional to $\frac{dR}{dt}\mathbf{n}$, being R the radius of each bacterium, see [3] for more details. In our case, $\mathbf{F}_{ext} = \mathbf{F}_i$ are interaction forces between bacteria, moving them as blocks. For spherical bacteria, we set $\mathbf{F}_i = \sum_{j=1}^{N} \mathbf{F}_{i,j} \delta_j$ with

$$\mathbf{F}_{i,j} = \begin{cases} \displaystyle\sum_{n=1, n \neq j}^{N} \frac{\sigma}{d_{min}} \mathbf{n}_{\mathrm{cm},n,j} & \text{if } d_{j,n} \leq d_{min}, \\ \displaystyle\sum_{n=1, n \neq j}^{N} \frac{\sigma \left(1 + \tanh\left(\frac{s_p - d_{j,n}}{v_p}\right)\right)}{2 d_{j,n}} \mathbf{n}_{\mathrm{cm},n,j} & \text{if } d_{j,n} > d_{min}, \end{cases} \tag{4}$$

where σ is a repulsion parameter, $\mathbf{n}_{\mathrm{cm},n,j} = \frac{\mathbf{X}_{c,j} - \mathbf{X}_{c,n}}{\|\mathbf{X}_{c,j} - \mathbf{X}_{c,n}\|}$ the unit vector that joins the centers of mass, and $d_{j,n}$ the distance between them. Here, δ_j equals 1 on the boundary $\mathbf{X}_j$ and vanishes on the rest. s_p controls at what distance the force begins to act and v_p its growth if the distance continues to decrease, see [3]. This force is easy to adjust and extend to rod-like shapes by tuning parameters [3], as opposed to the forces employed in [7]. Resorting to Morse potentials would be too expensive, whereas Lennard-Jones potentials seem too strong.

3 Dynamic Energy Budget Model for Cell Metabolism

The growth dynamics of the boundaries representing bacterial membranes is governed by bacterial metabolism. We use a Dynamic energy budget (DEB) [1, 9] model for each cell, informed by a set of relevant concentration fields.

Given N bacteria, their energy e_j and volume V_j, $j = 1, \ldots, N$, are governed by

$$\frac{de_j}{dt} = \nu' \left(\frac{S}{S + K_S} - e_j\right), \quad \frac{dV_j}{dt} = \left(r_j \frac{a_j}{a_M} - h_j\right) V_j, \quad r_j = \left(\frac{\nu' e_j - mg}{e_j + g}\right)^{+}, \tag{5}$$

where $\nu' = \nu e^{-\gamma \varepsilon}$, ν is the energy conductance, γ the environmental degradation coefficient, K_S a half-saturation coefficient, m the maintenance rate, g the investment ratio and a_M the target acclimation energy. The symbol $^{+}$ stands for 'positive part'. The factor r_j denotes the bacterial production rate. For 2D spherical bacteria

$V_j = \pi R_j^2$, and (5) implies

$$2\frac{dR_j}{dt} = \Big(r_j \frac{a_j}{a_M} - h_j\Big) R_j. \tag{6}$$

The aging q_j and hazard h_j variables represent damage on the cell, while a_j stands for acclimation, governed by

$$\frac{dq_j}{dt} = e_j(s_G \rho_x \frac{V_j}{V_T} q_j + h_a)(\nu - r_j) - (r_j + r_{e,j})q_j, \quad \frac{dh_j}{dt} = q_j - (r_j + r_{e,j})h_j, \tag{7}$$

$$\frac{dp_j}{dt} = -h_j p_j, \quad \frac{da_j}{dt} = (r_j + r_{e,j}) \left(1 - \frac{a_j}{a_M}\right)^+, \tag{8}$$

where ρ_x is the cell density, h_a the Weibull aging acceleration, s_G a multiplicative stress coefficient. Here, p_j is the probability of survival at time t. The factor $r_{e,j} = kr_j + k'$, for $k, k' > 0$ when the cell is and polymer (EPS) producer, otherwise it vanishes. The produced EPS is then $\frac{dV_{e,j}}{dt} = r_{e,j} V_j$. A fraction $\eta \in (0, 1)$ diffuses creating a concentration of monomers C_e, while the rest sticks to the bacteria. The limiting substrate concentration S and environmental degradation ε satisfy

$$\frac{dS}{dt} = -\nu' \frac{S}{S + K_S} \rho_x \sum_j \frac{V_j}{V_T} \delta_j + d_s \Delta S - \mathbf{u} \cdot \nabla S, \tag{9}$$

$$\frac{dC_e}{dt} = \eta \rho_x \sum_j r_{e,j} \frac{V_j}{V_T} \delta_j + d_e \Delta C_e - \mathbf{u} \cdot \nabla C_e, \tag{10}$$

$$\frac{d\varepsilon}{dt} = \nu_\varepsilon \rho_x \sum_j (r_j + \nu_m m) \frac{V_j}{V_T} \delta_j + d_\varepsilon \Delta \varepsilon - \mathbf{u} \cdot \nabla \varepsilon, \tag{11}$$

where ν_m is the maintenance respiratory coefficient, ν_ε is the environmental degradation coefficient and d_s, d_e, d_ε diffusion coefficients. Here V_T is a reference volume and $\delta_j = 1$ at cell j, it vanishes otherwise. We enforce no flux boundary conditions, except for S, which remains constant at the borders.

We couple the system of ordinary differential equations (5)–(7) and the reaction-diffusion equations (9)–(11) using a similar philosophy as that in IB models. We spread fields defined on bacteria using the cell volumes and rates as sources in equations (9)–(11). We interpolate global fields on the bacteria averaging values of S, ε in the region occupied by the cell. For each cell, the systems (5)–(7) are discretized order two Runge-Kutta schemes. The reaction-diffusion equations (9)–(11) are discretized by classical explicit finite difference schemes, first order in time and second order in space.

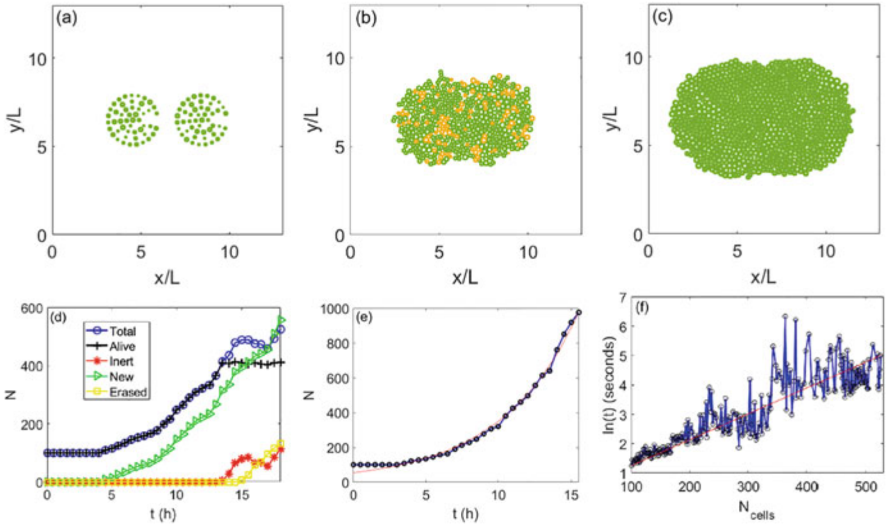


Fig. 1 (**a**) Initial configuration. Simulated configurations (**b**) after 18 hours with $h_0 = 0.4$ and (**c**) after 15 hours with $h_0 = 0$ (no death). While (**b**) has 412 live cells (green) and 113 dead cells (orange), (**c**) contains 920 live cells. Growth curve of cell types versus time (**d**) for simulation (**b**) and (**e**) for simulation (**c**). The red fit in (**e**) is $t \sim Ce^{\gamma N}$, where $C \sim 54.53$ and $\gamma \sim 0.1861$ [1/h], N being the number of cells. Panel (**f**) shows the times required to perform one computational step depending on the number of cells in simulation (**a**)–(**b**) (blue circles) compared to its exponential fit (red), $Ce^{\gamma t}$, where $C \sim 1.4421$ [s] and $\gamma \sim 0.0088$

4 Simulations of Biofilm Spread

For typical parameter values [1, 3], the IB and concentration submodels are quasistationary. Their solutions evolve as the immersed boundaries grow, shrink, divide or move due to interactions. We solve the DEB equations (5)–(7) in a time scale of hours, while updating the IB and concentration fields using time relaxation schemes to update the quasistationary fields. Results are displayed in Fig. 1. Cells $\mathbf{X}_j$ die when $1 - p_j > \frac{N_{init}}{N_a} + r\left(1 - \frac{N_{init}}{N_a}\right)$, N_a and N_{init} being the current and the initial number of bacteria while $r \in (0, 1)$ is a random number [3]. Cells divide in two when their size surpasses a critical value.

5 Conclusions and Perspectives

Modeling the behavior of cell aggregates such as bacterial biofilms confronts the difficulties of handling complicated interactions and geometries. We propose an immerse boundary approach with enhanced spatial resolution when compared to particle or cellular automata descriptions, since we can track individual deformations and fluid-structure interactions. This approach is computationally expensive if we aim to grow large clusters to see behaviors emanating at larger scales. However, High Performance Computing tools may help to overcome that burden. The present work focuses on spherical bacteria. Extensions to other shapes (rod-like, mixtures), geometries (interaction with barriers) and environments (inclusion of toxicants) can be envisaged [3]. In our current simulations the fluid flow has little relevance. Exploring interactions with the flow and its influence on the observed shapes [16] should be the subject of further work.

Acknowledgments Research partially supported by Spanish FEDER/MICINN-AEI grants MTM 2017-84446-C2-1-R and PID2020-112796RB-C21 (AC, RG), by fellowship PRE2018-083807 (RG), and by the MICINN 'Salvador de Madariaga" grant PRX18/00112 (AC). A.C. thanks C.S. Peskin for fruitful discussions and suggestions during a sabbatical stay at NYU.

References

1. B. Birnir, A. Carpio, E. Cebrián, P. Vidal, Dynamic energy budget approach to evaluate antibiotic effects on biofilms, Commun. Nonlinear Sci. Numer. Simulat. 54 (2002) 70–83.
2. A. Carpio, E. Cebrián, P. Vidal, Biofilms as poroelastic materials, Int. J. Non-Linear Mech. 109 (2019) 1–8.
3. A. Carpio, R. González-Albaladejo, Immersed boundary approach to biofilm spread on surfaces, Commun. Comput. Phys. 31 (1), 257–292, 2022.
4. R. Dillon, L. Fauci, A. Fogelson, D. Gaver, Modeling biofilm processes using the immersed boundary method, J. Comput. Phys. 129 (1996) 57–73.
5. R. Dillon, M. Owen, K. Painter, A single-cell-based model of multicellular growth using the immersed boundary method, In: Moving Interface Problems and Applications in Fluid Dynamics (pp. 1–16). (Contemporary Mathematics). American Mathematical Society, 2008.
6. H.C. Flemming, J. Wingender, The biofilm matrix, Nat. Rev. Microbiol. 8 (2010) 623–633.
7. M.A.A. Grant, B. Waclaw, R.J. Allen, P. Cicuta, The role of mechanical forces in the planar-to-bulk transition in growing *Escherichia coli* microcolonies, J. R. Soc. Interface 11 (2014) 20140400.
8. N. Høiby, T. Bjarnsholt, M. Givskov, S. Molin, O. Ciofu, Antibiotic resistance of bacterial biofilms, Int J Antimicrob Agents 35 (2010) 322–32.
9. T. Klanjscek, R.M. Nisbet, J.H. Priester, P.A. Holden, Modeling physiological processes that relate toxicant exposure and bacterial population dynamics, PLoS One 7(2) (2012) e26955.
10. L. A. Lardon, B. V. Merkey, S. Martins, et al, iDynoMiCS: next-generation individual-based modelling of biofilms, Environ. Microbiol. 13 (2011) 2416–34.
11. C. S. Laspidou, L. A. Spyrou, N. Aravas, B. E. Rittmann, Material modeling of biofilm mechanical properties, Math. Biosci. 251 (2014) 11–15.
12. Y. Li, A. Yun, J. Kim, An immersed boundary method for simulating a single axisymmetric cell growth and division, J. Math. Bio. 65 (2012) 653–675.
13. C.S. Peskin, D.M. McQueen, A general method for the computer simulation of biological systems interacting with fluids, Symposia of the Society for Experimental Biology 49 (1995) 265–76.
14. C.S. Peskin, The immersed boundary method, Acta Numerica 11 (2002) 479–517.
15. K.A. Rejniak, An immersed boundary framework for modelling the growth of individual cells: an application to the early tumour development, J. Theoret. Bio. 247 (2007) 186–204.
16. D. Rodriguez, B. Einarsson, A. Carpio, Biofilm growth on rugose surfaces, Phys. Rev. E 86 (2012) 061914.

17. A. Seminara, T.E. Angelini, J.N. Wilking, et al, Osmotic spreading of *Bacillus subtilis* biofilms driven by an extracellular matrix, Proc. Natl. Acad. Sci. USA 109 (2012) 1116–1121.
18. T. Storck, C. Picioreanu, B. Virdis, D.J. Batstone, Variable cell morphology approach for individual-based modeling of microbial communities, Biophys. J. 106 (2014) 2037–2048.
19. J.A. Stotsky, J.F. Hammond, L. Pavlovsky, et al, Variable viscosity and density biofilm simulations using an immersed boundary method, Part II: Experimental validation and the heterogeneous rheology-IBM, J. Comput. Phys. 317 (2016) 204–222.
20. R. Sudarsan, S. Ghosh, J.M. Stockie, H.J. Eberl, Simulating biofilm deformation and detachment with the immersed boundary method, Commun. Comput. Phys. 19 (2016) 682–732.

Numerical Simulation of a Four Serotype Dengue Fever Model

Gaby Folger and Kurt Chudej

Abstract Dengue fever is a vector-borne virus infection of the tropics and subtropics. It is transmitted by Asian tiger mosquitos and comes with four serotypes. These mosquitos are currently (re-)invading Europe, established Asian tiger mosquito populations are already known e.g. in a few German cities. Dengue fever cases were imported into Germany in pre-covid times usually by airtravel, but some autochthonous cases in Europe are already known. We present numerical simulations and optimal control results of a new four serotype dengue fever model including an imperfect vaccination of seropositive humans.

1 New Four Serotype Model of Dengue

We present a new four serotype dengue fever model [4], a generalization of the two-serotype dengue models [1–3, 5, 6]. Additionally the new model includes a vaccination of seropositive humans. This is consistent with current advice for available dengue vaccines.

The human population (subscript h) is subdivided into the following compartments: S_h susceptible humans, I_h^i with serotype i infected humans, S_h^i previously with serotype i infected humans which recovered and are now resistant to serotype i and are susceptible for serotypes $j \neq i$, V_h^i humans seropositive (and resistent) to serotype i and vaccinated (i.e. immune against serotypes $j \neq i$), I_h^{ij} humans which were previously infected with serotype i, recovered, and are now infected with serotype j, R_h humans which are resistant.

The vector population of (female) mosquitoes (subscript v), is subdivided into the compartments: S_v susceptible adult mosquitoes, I_v^i adult mosquitoes infected

G. Folger · K. Chudej (✉)
Universität Bayreuth, Lehrstuhl für Wissenschaftliches Rechnen, Bayreuth, Germany
e-mail: gaby.folger@uni-bayreuth.de; kurt.chudej@uni-bayreuth.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_9

with serotype i. For simplicity, we assume that a vector cannot carry more than one serotype of the dengue virus.

The following general assumptions hold: Both vectors and humans are neither born infected nor resistant. The size of the human population is constant at any time. Neither emigration nor immigration are considered. The passing of a certain percentage of the population is modeled through the proportionality factor μ_h, with $1/\mu_h$ denoting an average lifespan. The same holds for the adult mosquito population with factor μ_v. Humans recover from the disease (of any serotype) at a rate η_h. The infection rates are given by $b_{i,h}$, $b_{i,v}$. The infection rates of a second infection are modified by a factor δ_i. Depending on the vector they come into contact with, humans either change to the compartments I_h^i or I_h^{ij}.

This yields the following 31 ODEs (neglecting third and fourth infections in the model)

$$\begin{aligned}
\frac{dS_h}{dt} &= \mu_h - \Big(\mu_h + \sum_{k=1}^{4} b_{k,h} I_v^k\Big) S_h \\
\frac{dI_h^i}{dt} &= b_{i,h} I_v^i S_h - (\eta_h + \mu_h) I_h^i \\
\frac{dS_h^i}{dt} &= \eta_h I_h^i + \theta V_h^i - \Big(\psi + \mu_h + \sum_{l=1, l\neq i}^{4} \delta_i b_{l,h} I_v^l\Big) S_h^i \\
\frac{dV_h^i}{dt} &= \psi S_h^i - \Big(\theta + \mu_h + \sigma \sum_{l=1, l\neq i}^{4} \delta_i b_{l,h} I_v^l\Big) V_h^i \\
\frac{dI_h^{ij}}{dt} &= \delta_i b_{j,h} I_v^j (S_h^i + \sigma V_h^i) - (\mu_h + \eta_h) I_h^{ij} \\
\frac{dR_h}{dt} &= \eta_h \Big(\sum_{k=1}^{4} \sum_{l=1, l\neq k}^{4} I_h^{kl}\Big) - \mu_h R_h \\
\frac{dS_v}{dt} &= \mu_v - \Big[\mu_v + \sum_{k=1}^{4} b_{k,v} \Big(I_h^k + \sum_{l=1, l\neq k}^{4} I_h^{lk}\Big)\Big] S_v \\
\frac{dI_v^i}{dt} &= b_{i,v} \Big(I_h^i + \sum_{l=1, l\neq i}^{4} I_h^{li}\Big) S_v - \mu_v I_v^i
\end{aligned} \tag{1}$$

together with the positive invariant set $\Omega = \Big\{(S_h, I_h^i, S_h^i, V_h^i, I_h^{ij}, R_h, S_v, I_v^i) \in \mathbb{R}^{31}_{\geq 0} |$ $S_h + R_h + \sum_{i=1}^{4}\Big(I_h^i + S_h^i + V_h^i + \sum_{j=1, j\neq i}^{4} I_h^{ij}\Big) \leq 1,\ S_v + \sum_{i=1}^{4} I_v^i \leq 1\Big\}$. The basic reproduction number is given by $\mathcal{R}_0 = \max_i \mathcal{R}_{0i}$, $\mathcal{R}_{0i} = \sqrt{\frac{b_{i,h} b_{i,v}}{\mu_v(\eta_h + \mu_h)}}$ and the invasion numbers are calculated as

$$\mathcal{R}_{ij} = \frac{\mathcal{R}_{0j}}{\mathcal{R}_{0i}} \sqrt{1 + \frac{\delta_i \mu_v \eta_h}{b_{i,v}\mu_h + \mu_v(\eta_h + \mu_h)} (\mathcal{R}_{0i}^2 - 1)}.$$

2 Numerical Simulation

Tables 1 and 2 list the used parameter values and asymmetric infection rates. We present some numerical simulation results for the model without vaccination. Therefore $\psi \equiv 0$ and $V^i(0) = 0$. The used assumption for the infection rates yields different scenarios which are not too far apart. Inserting the values, all the partial basic reproduction numbers $\mathcal{R}_{0i}$ and the invasion numbers $\mathcal{R}_{ij}$ are above one, such that an epidemic can be expected. The following initial values are used, all other initial values are zero:

$$S_h(0) = 0.99952,\ S_v(0) = 0.05,$$

$$I_h^1(0) = 0.0002,\ I_h^2(0) = 0.0001,\ I_h^3(0) = 0.00017,\ I_h^4(0) = 0.00001$$

By this choice of initial values an additional asymmetry of the system is investigated. Figures 1 and 2 show the development in the compartments of the

Table 1 Parameters and infection rates for modelling

Parameter	Value		Meaning
μ_h^{-1}	$65 \cdot 365$	[day]	Average life span of humans
η_h^{-1}	6	[day]	Average disease duration of humans
μ_v^{-1}	21	[day]	Average life span of mosquitos
δ_i	1	[–]	Influence of a previous disease
$b_{1,h}$	0.21	$[\frac{1}{\text{day}}]$	Infection rate (serotype 1) vector to human
$b_{2,h}$	$b_{1,h} \cdot \epsilon$	$[\frac{1}{\text{day}}]$	Infection rate (serotype 2) vector to human
$b_{3,h}$	$b_{1,h} \cdot 0.95$	$[\frac{1}{\text{day}}]$	Infection rate (serotype 3) vector to human
$b_{4,h}$	$b_{1,h} \cdot 0.9$	$[\frac{1}{\text{day}}]$	Infection rate (serotype 4) vector to human
$b_{i,v}$	$b_{i,h} \cdot 1.6$	$[\frac{1}{\text{day}}]$	Infection rate (serotype i) human to vector
ϵ	0.9 or 1.1	[–]	Scaling parameter for different scenarios
ψ	0 or control	$[\frac{1}{\text{day}}]$	Vaccination rate
θ	$1/(0.7 \cdot 365)$	$[\frac{1}{\text{day}}]$	Rate at which the vaccination loses effect
σ	0.02	[–]	High efficiency of the vaccination

Table 2 Scenarios of infection coefficents

Scenario	Value	Ordering of infection coefficients
A	$\epsilon = 0.9$	$b_{1,h} > b_{3,h} > b_{4,h} = b_{2,h}$
B	$\epsilon = 1.1$	$b_{2,h} > b_{1,h} > b_{3,h} > b_{4,h}$

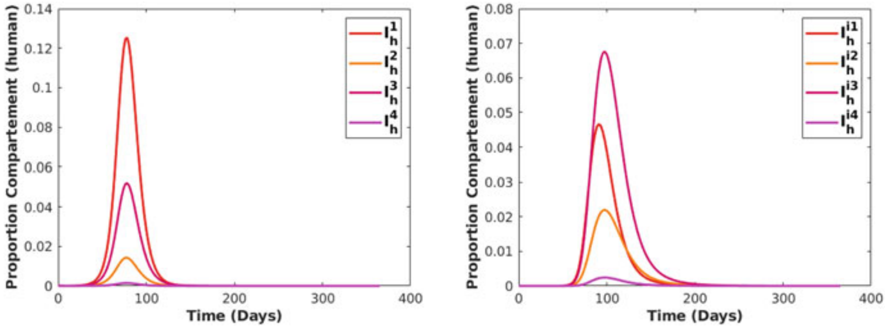


Fig. 1 Infected humans for $\epsilon = 0.9$

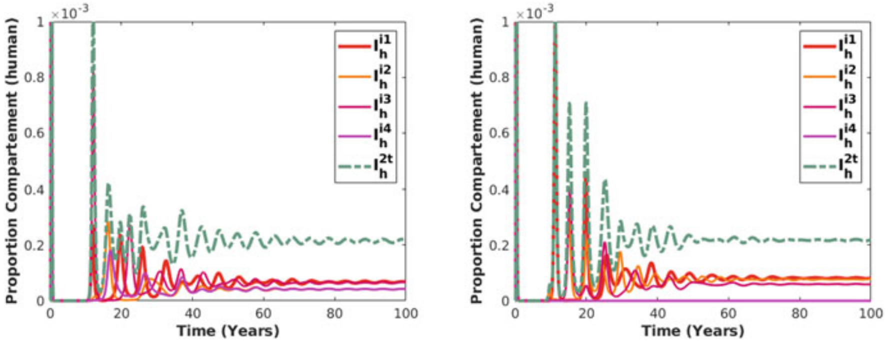


Fig. 2 Second infections for $\epsilon = 0.9$ (left) and $\epsilon = 1.1$ (right)

first infections and the second infections. In the legend of the figures I^{i1} is an abbreviation for $\sum_i I^{i1}$ and so forth. Figure 1 shows the expected strong outbreak in the infected compartment with the highest infection rate as well as lower peaks in the compartments of serotype 3 and 4. Due to the permanent resistance to the first infection the highest peak appears for the second infection for the serotype of the second largest infection rate (Fig.1 right). The second largest peak appears for the second infection for the serotype with the largest infection rate. If one considers a longer time horizon of 100 years, interesting things happen: For sufficiently different infection rates, one or two serotypes are eradicated (Fig. 2). This phenomenon indicates multiple equilibria whose stability depends on the ratios between the infection rates $b_{h,i}$. If the infection rates are close enough, all serotypes coexist. On the other hand, the total number of second infections $I_h^{2t} := \sum_{i,j} I_h^{ij}$ approaches approximately identical values over this long time horizon.

3 Optimal Control

For optimal control the previous model is enhanced with a mosquito control c_v (e.g., spraying adulticide). Therefore μ_v is substituted by $\mu_v + c_v$ in the ODE.

The following cost functional is minimized, see also Table 3:

$$J = \int_0^{t_f} \left[\gamma_1 \cdot \left(\sum_{i=1}^{4} I_h^i\right)^2 + \gamma_2 \cdot \left(\sum_{i=1}^{4} \sum_{j=1, j\neq i}^{4} I_h^{ij}\right)^2 + \gamma_3 \cdot c_v^2 + \gamma_4 \cdot \psi^2\right] dt$$

Figures 3 and 4 show the optimal controls (vaccination rate ψ and mosquito control c_v) and the trajectories of the humans, which are infected for a second time, for case A–D.

Mosquito control starts directly at the initial time and reaches its peak with the peak period of infected humans and levels off afterwards. Since only seropositive humans can be vaccinated, vaccination starts later. Significantly lower infection numbers are observed in the optimal control case. The peak levels vary depending on the weighting.

Table 3 Overview of the different cases with their weights for $\epsilon = 0.9$

	Weights	Costs
Case A	$\gamma_1 = 0.17, \gamma_2 = 0.17, \gamma_3 = 0.33, \gamma_4 = 0.33$	0.1744
Case B	$\gamma_1 = 0.15, \gamma_2 = 0.15, \gamma_3 = 0.55, \gamma_4 = 0.15$	0.1503
Case C	$\gamma_1 = 0.15, \gamma_2 = 0.15, \gamma_3 = 0.15, \gamma_4 = 0.55$	0.1529
Case D	$\gamma_1 = 0.15, \gamma_2 = 0.55, \gamma_3 = 0.15, \gamma_4 = 0.15$	0.1793

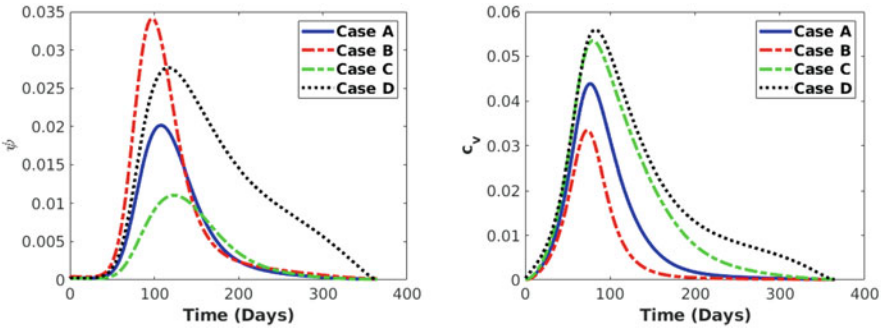


Fig. 3 Optimal controls for $\epsilon = 0.9$

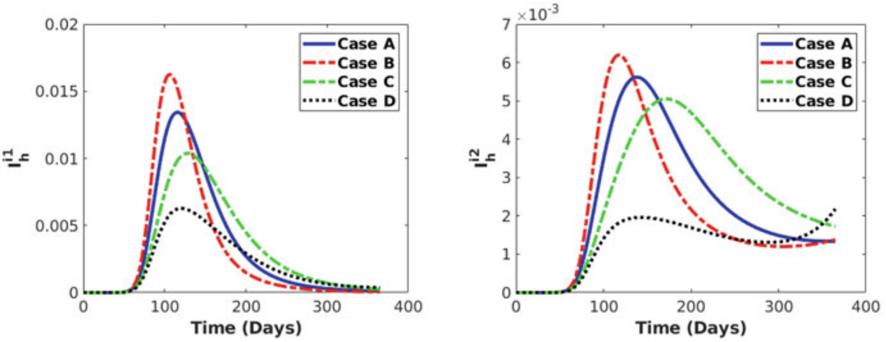


Fig. 4 Second infections for $\epsilon = 0.9$

4 Conclusion

A modeling with four serotypes is possible in principle, but calculations can only be handled numerically due to the sharp increase in complexity. A very interesting numerical result is that in the case of asymmetric infection rates, a competitive behavior of the serotypes can be observed. This can lead to the extinction of up to two serotypes.

References

1. Chudej, K., Albrecht, G., Jende, L.: Vereinfachung eines Denguefieber-Modells mit zwei Serotypen und einer extrinsischen Inkubationszeit der Vektoren. In: Wittmann, J. (ed.) Simulation in Umwelt- und Geowissenschaften, pp. 217–227. Shaker, Düren (2019)
2. Chudej, K., Fischer, A.: Optimal vaccination strategies for a new dengue model with two serotypes. IFAC PapersOnLine **51-2**, 13–18 (2018)
3. Esteva, L., Vargas, C.: Coexistence of different serotypes of dengue virus. Journal of mathematical biology **46-1** 31–47 (2003).
4. Folger, G.: Modellierung, Analyse und Optimale Steuerung von gefährlichen Krankheiten. PhD, Universität Bayreuth, 2021.
5. Herath, M., Albrecht, G., Chudej, K.: Ein asymmetrisches zwei Serotyp Dengue Fieber Modell mit Kontrollmaßnahmen. In: Wittmann, J. (ed.) Simulation in den Umwelt- und Geowissenschaften, pp. 191–202. Shaker, Düren (2020)
6. Zheng, T.-T., Nie, L.-F.: Modelling the transmission dynamics of two-strain Dengue in the presence awareness and vector control. Journal of theoretical biology **443**, 82–91 (2018)

The Effect of the Number of Neural Networks on Deep Learning Schemes for Solving High Dimensional Nonlinear Backward Stochastic Differential Equations

Lorenc Kapllani

Abstract We consider the deep learning based scheme proposed in [W. E and J. Han and A. Jentzen, Commun. Math. Stat., 5 (2017), pp. 349–380] and study the effect of the number of neural networks on the gradient of the solution. We demonstrate that using one neural network improves its numerical stability for the whole path and also reduces the computational time. This is illustrated with several 100-dimensional nonlinear backward stochastic differential equations including nonlinear pricing problems in finance.

1 Introduction

In this work we consider the high dimensional *forward backward stochastic differential equation (FBSDE)* of the form

$$\begin{cases} dX_t = \mu(t, X_t)\,dt + \sigma(t, X_t)\,dW_t, \quad X_0 = x_0, \\ -dY_t = f(t, X_t, Y_t, Z_t)\,dt - Z_t\,dW_t, \\ Y_T = \xi = g(X_T), \end{cases} \tag{1}$$

where $X_t, \mu \in \mathbb{R}^n$, σ is a $n \times d$ matrix, $W_t = \left(W_t^1, \ldots, W_t^d\right)^\top$ is a d-dimensional Brownian motion, $f(t, X_t, Y_t, Z_t) : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d} \to \mathbb{R}^m$ is the driver function and ξ is the terminal condition. The existence and uniqueness of the solution of (1) are proven in [10]. In the sequel of this work, we investigate the effect of the number of neural networks in [4] that solve (1).

In the recent years, many numerical methods have been proposed for solving BSDEs, e.g., [1, 11, 15], which are not suitable for high-dimensional problems

L. Kapllani (✉)
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: kapllani@math.uni-wuppertal.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_10

as in physics or finance [9] (also sparse-grids [14] or parallel computing [6, 8]) due to exponential increase of algorithm complexity. Recently machine learning schemes show that they can deal with high dimensions for reasonable computational time [4, 5, 12, 13]. We study the well known deep learning based algorithm in [4] (we refer to it as SDNN-approach in the rest of the paper, where SDNN stands for Stacked Deep Neural Networks) where the gradient of the solution (process Z) is approximated by fully-connected neural networks. The authors in [3] analyzed it using different architectures. However, no result was shown regarding the instability of Z for SDNN-approach due to the use of different deep neural networks at each time layer. To have a more numerically stable algorithm, we study the effect of reducing the number of neural networks. In the sequel, we refer this scheme as DNN-approach.

The outline of the paper is organized as follows. In the next section, we describe the SDNN- and DNN-approaches. In Sect. 3, we illustrate our findings with several numerical tests. Section 4 concludes this work.

2 The SDNN-Approach and DNN-Approach

The Feynman-Kac formula and forward discretization of FBSDE are needed to formulate the FBSDE as a learning problem. Let us consider that the terminal value Y_T is of the form $g(X_T^{t,x})$, where $X_T^{t,x}$ denotes the solution of forward SDE in (1) starting from x at time t. Then, the solution $(Y_t^{t,x}, Z_t^{t,x})$ of (1) can be presented as [9]

$$Y_t^{t,x} = u(t,x), \quad Z_t^{t,x} = \big(\nabla u(t,x)\big)\sigma(t,x) \quad \forall t \in [0,T), \tag{2}$$

where $u(t,x)$ is the solution of the following semi-linear parabolic PDE:

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{n} \mu_i(t,x)\frac{\partial u}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{n} (\sigma\sigma^\top)_{i,j}(t,x)\frac{\partial^2 u}{\partial x_i x_j} + f\big(t,x,u,(\nabla u)\sigma\big) = 0,$$

with $u(T,x) = g(x)$. This is the Feynman-Kac formula. Using

$$\Delta = \{t_i | t_i \in [0,T], i = 0, 1, \cdots, N, t_i < t_{i+1}, \Delta t = t_{i+1} - t_i, t_0 = 0, t_N = T\}$$

and the notation $X_i = X_{t_i}$, $W_i = W_{t_i}$, $\Delta W_i = W_{i+1} - W_i$ and the approximated process as $X_i^\Delta = X_{t_i}^\Delta$, the discretization of (1) using the well-known Euler scheme is

$$X_{i+1}^\Delta = X_i^\Delta + \mu\left(t_i, X_i^\Delta\right)\Delta t + \sigma\left(t_i, X_i^\Delta\right)\Delta W_i,$$

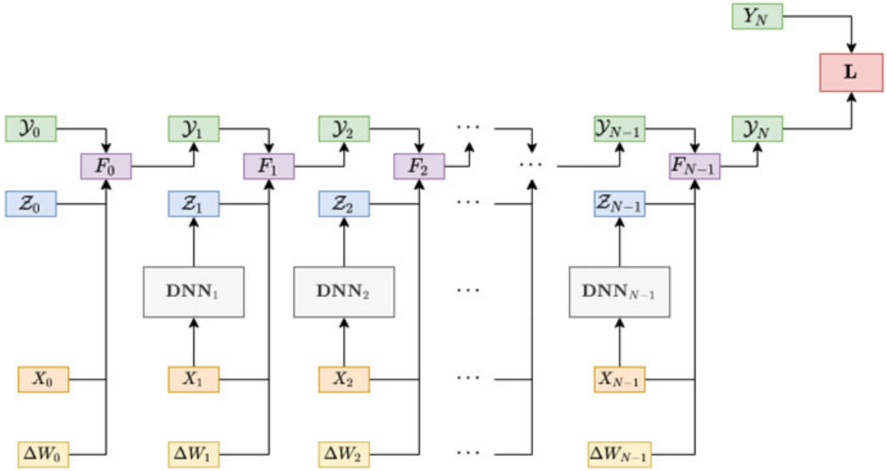


Fig. 1 Graph of the SDNN-approach

and

$$
\begin{aligned}
Y_{i+1}^{\Delta} &= Y_i^{\Delta} - f\left(t_i, X_i^{\Delta}, Y_i^{\Delta}, Z_i^{\Delta}\right)\Delta t + Z_i^{\Delta}\Delta W_i, \\
&:= F(t_i, X_i^{\Delta}, Y_i^{\Delta}, Z_i^{\Delta}, \Delta t, \Delta W_i).
\end{aligned} \tag{3}
$$

where $i = 0, 1, \dots, N-1$ and $\Delta W_i \sim \mathcal{N}(0,\ \Delta t)$.

The numerical approximation of (Y^{Δ}, Z^{Δ}) in the SDNN-approach (Fig. 1) is designed as follows: starting from an estimation $(\mathcal{Y}_0(\theta), \mathcal{Z}_0(\theta))$ of $(Y_0^{\Delta}, Z_0^{\Delta})$, and then using at each time step t_i, $i = 1, 2, \dots, N-1$ a different feedforward deep neural network $\psi^{\varrho}_{i,k,L}(x;\theta)\colon \mathbb{R}^d \to \mathbb{R}^{1\times d}$ to approximate Z_i^{Δ} as $\mathcal{Z}_i(\theta)$ and Y_i^{Δ}, $i = 1, 2, \dots, N$ as $\mathcal{Y}_i(\theta)$ with (3), where the output $\mathcal{Y}_N(\theta)$ aims to match the terminal condition $g(X_T^{\Delta})$ of the BSDE (using Adam gradient descent-type optimizer with mini-batches):

$$
\mathbb{E}\left[|g(X_T^{\Delta}) - \mathcal{Y}_N(\theta)|^2\right].
$$

Note that i represents the i-th network, k is the number of neurons, L are the number of hidden layers, ϱ is the activation function, the input x of the network is the Markovian process X_i^{Δ} and θ are network parameters. Specifically, the networks have 4 global layers, where hidden layers have d+10 neurons, the rectifier function $\mathbb{R} \ni x \to \max\{0, x\} \in [0, \infty)$ is used as the activation function, the weights are initialized using a normal or a uniform distribution and batch normalization is also used. For the DNN-approach, we consider $p < N-1$ networks, i.e. 1 network for consecutive subintervals, with input x having the time discretization t_i (to handle non-stationarities) and the Markovian process X_i^{Δ} (due to Feynman-Kac formula), with 6 global layers (2 hidden layers more than SDNN-approach for

Table 1 The dimension of the parameters

SDNN-approach	$d+1+(N-1)(2d(d+10)+(d+10)^2+4(d+10)+2d)$
DNN-approach	$p\left(2d+1+(2d+5)(d+10)+3(d+10)^2\right)$

a better accuracy). For sufficiently regular solutions, the gradient between two time points should be close. Therefore, the numerical stability of Z is affected from the number of DNNs. The dimension of the parameters $\rho \in \mathbb{N}$ for both approaches are given in Table 1. The complexity in the DNN-approach is lower for less number of networks, namely p.

3 Numerical Results

In this section we study the DNN-approach by comparing it to the SDNN-approach in several high dimensional examples. The results are presented using 10 independent runs with Tensorflow 1.15 from Google Colab. We start with an example with analytical solution where the driver function depends on Y and Z.

Example 1 Consider the Burgers type FBSDE [4]

$$\begin{cases} dX_t = \sigma\, dW_t, \quad X_0 = 0, \\ -dY_t = \left(Y_t - \frac{2+d}{2d}\right)\left(\sum_{i=1}^d Z_t^i\right) dt - Z_t\, dW_t, \\ Y_T = 1 - \frac{1}{1+\exp\left(T + \frac{1}{d}\sum_{i=1}^d X_T^i\right)}, \end{cases}$$

where $W_t = (W_t^1, W_t^2, \cdots, W_t^d)^\top$, $X_t = (X_t^1, X_t^2, \cdots, X_t^d)^\top$, $Z_t = (Z_t^1, Z_t^2, \cdots, Z_t^d)$. The exact solution is $(Y_0, Z_0) \doteq (0.5, (0.1768, \cdots, 0.1768))$ with $d = 50$, $T = 0.2$ and $\sigma = \frac{d}{\sqrt{2}}$. We consider the same hyperparameters for both the SDNN- and DNN-approach, where the learning rate is $1e{-}2$, 8000 optimization iterations, 256 validation sample and a batch size of 64, which are used also for next examples if not specified. The authors in [4] used different hyperparameters and discretization values. The results are reported in Table 2 for $N = 40$. Note that $p = N - 1$ represents the SDNN-approach, $|\cdot|$ is the absolute value and $s(\cdot)$ represents the standard deviation. Moreover, $\epsilon_{Y_0} = |Y_0 - \mathcal{Y}_0|$, $\mathbf{Z}_0 = \frac{1}{d}\sum_{i=1}^d \mathcal{Z}_0^i$ and $\epsilon_{Z_0} = \frac{\sum_{i=1}^d |Z_0^i - \mathcal{Z}_0^i|}{d}$.

From Table 2 we observe that the DNN-approach with one network gives higher accuracy for both processes Y and Z, for less computation time. Increasing the number of networks worsens the performance. To illustrate how good paths of each process are approximated, we display the averages of paths for Y as $\bar{Y}$ and Z as $\bar{Z}$, and the averages of approximated paths for $\mathcal{Y}$ as $\bar{\mathcal{Y}}$ and $\mathcal{Z}$ as $\bar{\mathcal{Z}}$ in Fig. 2, where the average over the dimension is also considered for the Z process, in order to have one value at each time point.

Table 2 The results for Example 1

p	$\mathcal{Y}_0$	ϵ_{Y_0}	$s(\epsilon_{Y_0})$	$\mathbf{Z}_0$	ϵ_{Z_0}	$s(\epsilon_{Z_0})$	Time
1	0.5196	0.0350	0.0198	0.1983	0.1250	0.0909	743.36
2	0.5409	0.1021	0.0983	0.3273	0.1814	0.1399	798.12
4	0.5312	0.1255	0.0651	0.5721	0.5092	0.2578	791.48
N − 1	1.2425	0.7425	0.0552	2.8510	2.6743	0.0290	1721.33

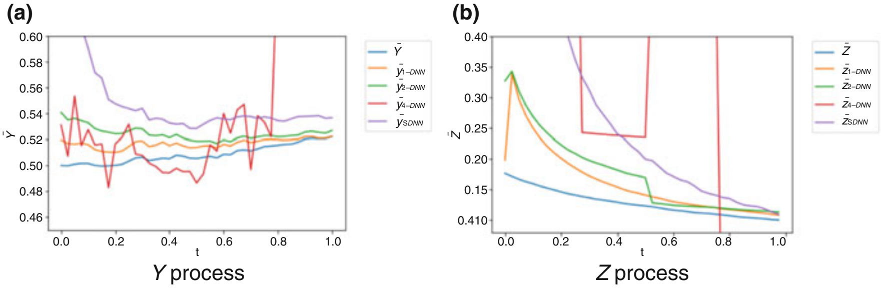


Fig. 2 The comparison of averages of the exact path $\bar{Y}$, $\bar{Z}$ and the approximated paths $\bar{\mathcal{Y}}$, $\bar{\mathcal{Z}}$ for Example 1

For a better view of the approximation of the whole path, we limit the axis for Y and Z, since some results are far from the exact solution. We see that the DNN-approach with one network approximates paths of both processes much better in this example when $d = 50$. Next, we consider an example with a driver function where the Z process grows quadratically.

Example 2 Consider the nonlinear BSDE [7]

$$\begin{cases} -dY_t = \left(\|Z_t\|^2_{\mathbb{R}^{1\times d}} - \|\nabla\psi(t, W_t)\|^2_{\mathbb{R}^d} - \left(\partial_t + \frac{1}{2}\Delta\right)\psi(t, W_t)\right) dt - Z_t\, dW_t, \\ Y_T = \sin\left(\|W_T\|^{2\alpha}_{\mathbb{R}^d}\right), \end{cases}$$

where $\psi(t, W_t) = \sin\left(\left(T - t + \|W_t\|^2_{\mathbb{R}^d}\right)^{\alpha}\right)$. The exact solution is $(Y_0, Z_0) \doteq (0.8415, (0, \cdots, 0))$ with $d = 100$, $T = 1$ and $\alpha = 0.4$. Here we choose $d = 100$ to compare both the approaches in a higher dimension. We set the optimization iterations to $m = 4000$, and report the results in Table 3 with $N = 40$. We observe that the DNN-approach with one network gives again better results for both processes, even for the whole path displayed in Fig. 3.

Finally we consider an example without analytical solution in the case of $d = 100$, the problem of option pricing with different interest rates, also studied in [4, 5, 12].

Example 3 Consider the different interest rates option pricing FBSDE [2]

Table 3 The results for Example 2

p	$\mathcal{Y}_0$	ϵ_{Y_0}	$s(\epsilon_{Y_0})$	$\mathbf{Z}_0$	ϵ_{Z_0}	$s(\epsilon_{Z_0})$	Time
1	0.8583	0.0168	0.0077	0.0000	0.0087	0.0027	1197.47
2	0.8813	0.0398	0.0146	−0.0001	0.0120	0.0027	1161.98
4	0.9640	0.1226	0.0199	−0.0005	0.0188	0.0021	1304.86
N − 1	1.2541	0.4126	0.0247	−0.0004	0.0381	0.0037	1493.97

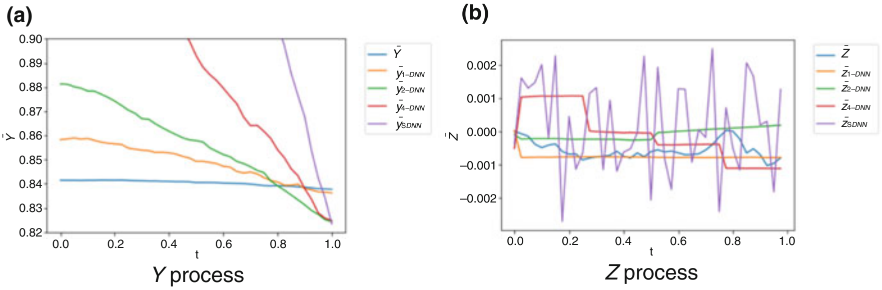


Fig. 3 The comparison of averages of the exact path $\bar{Y}$, $\bar{Z}$ and the approximated paths $\bar{\mathcal{Y}}$, $\bar{\mathcal{Z}}$ for Example 2

Table 4 The results for Example 3

p	$\mathcal{Y}_0$	$\lvert Y_0 - \mathcal{Y}_0 \rvert$	$s(\lvert Y_0 - \mathcal{Y}_0 \rvert)$	Time
1	21.0906	0.2082	0.0476	1137.93
2	21.0626	0.2362	0.0971	1144.90
4	21.0284	0.2704	0.1399	1206.71
N − 1	21.1140	0.1848	0.1017	1845.27

$$\begin{cases} dS_t &= \mu S_t\,dt + \sigma S_t\,dW_t, \qquad S_0 = S_0, \\ -dY_t &= -R^l Y_t - \frac{\mu - R^l}{\sigma} \sum_{i=1}^{d} Z_t^i + \left(R^b - R^l\right) \max\left(\frac{1}{\sigma} \sum_{i=1}^{d} Z_t^i - Y_t, 0\right) dt - Z_t\,dW_t, \\ Y_T &= \max\big(\max_{d=1,\cdots,D}(S_{T,d} - K_1, 0\big) - 2\max\big(\max_{d=1,\cdots,D}(S_{T,d} - K_2, 0\big), \end{cases}$$

where $S_t = (S_t^1, S_t^2, \cdots, S_t^d)^\top$. The benchmark value with $T = 0.5$, $\mu = 0.06$, $\sigma = 0.2$, $R^l = 0.04$, $R^b = 0.06$, $K_1 = 120$, $K_2 = 150$ and $S_0 = 100$ is $Y_0 \doteq 21.2988$ [5]. Using a learning rate of $5e-2$ and 4000 optimization iterations, we present the results in Table 4 for $N = 40$, which shows comparable results for DNN-approach with one network and SDNN-approach.

4 Conclusions

In this work we have proposed the DNN-approach to improve the deep learning scheme [4]. With our numerical analysis we demonstrate that the DNN-approach with one neural network can give comparable approximation for Y and better approximation for Z on the whole time domain for lower computational cost.

References

1. Bender, C., Zhang, J.: Time discretization and Markovian iteration for coupled FBSDEs. Ann. Appl. Probab. **18**(1), 143–177 (2008).
2. Bergman, Y.Z.: Option pricing with differential interest rates, Rev. Finan. Stud. **8**(2), 475–500 (1995).
3. Chan-Wai-Nam, Q., Mikael, J., Warin, X.: Machine learning for semi linear PDEs, J. Sci. Comput. **79**(3), 1667–1712 (2019).
4. E, W., Han, J., Jentzen, A.: Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations, Commun. Math. Stat. **5**(4), 349–380 (2017).
5. E, W., Hutzenthaler, M., Jentzen, A., Kruse, T.: On multilevel picard numerical approximations for high-dimensional nonlinear parabolic partial differential equations and high-dimensional nonlinear backward stochastic differential equations, J. Sci. Comput. **79**(3), 1534–1571 (2019).
6. Gobet, E., López-Salas, J.G., Turkedjiev, P., Vázquez, C.: Stratified regression Monte-Carlo scheme for semilinear PDEs and BSDEs with large scale parallelization on GPUs, SIAM J. Sci. Comput. **38**(6), C652–C677 (2016).
7. Gobet, E., Turkedjiev, P.: Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions, Math. Comp. **85**(299), 1359–1391 (2015).
8. Kapllani, L., Teng, L.: Multistep schemes for solving backward stochastic differential equations on GPU. arXiv preprint arXiv:1909.13560 (2019).
9. Karoui, N.E., Peng, S., Quenez, M.C.: Backward stochastic differential equations in finance. Math. Finan. **7**(1), 1–71 (1997).
10. Pardoux, E., Peng, S.: Adapted solution of a backward stochastic differential equation, Syst. Control. Lett. **14**(1), 55–61 (1990).
11. Ruijter, M.J., Oosterlee, C.W.: A Fourier cosine method for an efficient computation of solutions to BSDEs, SIAM J. Sci. Comput. **37**(2), A859–A889 (2015).
12. Teng, L.: A review of tree-based approaches to solve forward-backward stochastic differential equations, arXiv preprint arXiv:1809.00325v4 (2019).
13. Teng, L.: Gradient boosting-based numerical methods for high-dimensional backward stochastic differential equations, arXiv preprint arXiv:2107.06673 (2021).
14. Zhang, G.: A sparse-grid method for multi-dimensional backward stochastic differential equations, J. Comput. Math. **31**(3), 221–248 (2013).
15. Zhao, W., Zhang, G., Ju, L.: A stable multistep scheme for solving backward stochastic differential equations, SIAM J. Numer. Anal. **48**(4), 1369–1394 (2010).

Qualitatively Correct Numerical Methods for the Basic Ross–Macdonald Malaria Model

István Faragó, Miklós E. Mincsovics, and Rahele Mosleh

Abstract We investigate the qualitative performance of different numerical methods applied to the Ross-Macdonald malaria model. It is known that for this model a certain set is positively invariant and the question is that the discrete system which is obtained from the model by the application of a numerical method possesses this property or not. This property called dynamical consistency is the objective of this study. We consider a method qualitatively correct if the resulted discrete system inherits this property. We investigate the explicit and implicit Euler methods, the latter also with Newton iteration as a sub-procedure, moreover a non-local discretization method and finally, the explicit Euler method combined with step-size functions.

I. Faragó
Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary

Department of Applied Analysis and Computational Mathematics, Eötvös Loránd University, Budapest, Hungary

Large Networks Research Group, ELKH, Budapest, Hungary
e-mail: faragois@caesar.elte.hu

M. E. Mincsovics
Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary

Large Networks Research Group, ELKH, Budapest, Hungary
e-mail: mincso@cs.elte.hu

R. Mosleh (✉)
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: rmosleh@math.bme.hu

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_11

1 The Ross-Macdonald Model of Malaria Propagation

A diversity of mathematical models have been propounded to study malaria transmission. The basic Ross-Macdonald model is the most preliminary brought up by Ross and later modified by Macdonald.

This model reads as

$$\begin{cases} \dot{x}(t) = \alpha y(t)(1 - x(t)) - rx(t) \\ \dot{y}(t) = \beta x(t)(1 - y(t)) - \mu y(t). \end{cases} \tag{1}$$

where $x(t)$ and $y(t)$ represent densities of the infected humans and mosquitoes at time $t \geq 0$. Clearly, the minimal requirement for this model is that these quantities have to behave like densities, e.g. $x(t) \in [0, 1]$ and $y(t) \in [0, 1]$ hold for $t > 0$, provided that $x(0) \in [0, 1]$ and $y(0) \in [0, 1]$. We call this property *Density Preservation Property*, shortly DPP. This is indeed true for this model, see e.g. [2]. The next question is that if we apply some numerical method to approximate the solution of (1), then will the discrete version of the DPP—which we call shortly *DDPP* be valid for the discrete model or not?

2 Numerical Solution of the Ross-Macdonald Model

To approximate the solution of this model we apply the prototypes of the explicit and the implicit methods, the explicit and the implicit Euler method. We will analyze their performance from qualitative point of view.

By applying the explicit Euler method—shortly EEM—to the Ross-Macdonald model we attain

$$\begin{cases} \dfrac{x_{n+1} - x_n}{\Delta t} = \alpha y_n(1 - x_n) - rx_n \\ \dfrac{y_{n+1} - y_n}{\Delta t} = \beta x_n(1 - y_n) - \mu y_n. \end{cases} \tag{2}$$

It is known, see [2], that the DDPP is valid for the discrete model (2), if the step size $\Delta t \in (0, h^*]$, where

$$h^* = \min\left\{\frac{1}{r}, \frac{1}{\mu}, \frac{1}{\alpha}, \frac{1}{\beta}\right\}. \tag{3}$$

Since the above bound is sharp, this means that the DDPP is not valid for any step size Δt, there is a restriction, namely the step size have to be small enough.

By applying the implicit Euler method—shortly IEM—to the Ross-Macdonald model we attain

$$\begin{cases} \dfrac{x_{n+1}-x_n}{\Delta t} = \alpha y_{n+1}(1-x_{n+1}) - r x_{n+1} \\ \dfrac{y_{n+1}-y_n}{\Delta t} = \beta x_{n+1}(1-y_{n+1}) - \mu y_{n+1}. \end{cases} \tag{4}$$

It is proved that system (4) possesses the DDPP unconditionally, which means that there is no restriction for the step size Δt, see [2]. Consequently, the implicit Euler method is much better from this point of view than its explicit twin. However, we have to pay a price for it. Equation (4) is an implicit system of equations for x_{n+1}, y_{n+1} which needs to be solved. Fortunately, the Ross–Macdonald model is simple in the sense that we are able to directly express the unknowns and with that we get an explicit formula.

Since usually this cannot be expected, it is worth to explore some method which works generally. An implicit method results in a discrete model which solution requires a sub-procedure at each step to approximate the solution at the new time-level and the most popular method to do this is the Newton-iteration. In our case this means

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k+1} = \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} - \left[f'\left(\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} \right) \right]^{-1} f\left(\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} \right), \tag{5}$$

where k is the Newton-iteration step counter and

$$\left[f'\left(\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} \right) \right]^{-1} f\left(\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} \right) = \frac{1}{\det} \begin{pmatrix} 1+\Delta t\left(\beta x_{n+1}^k + \mu\right) & -\Delta t\alpha\left(x_{n+1}^k - 1\right) \\ -\Delta t\beta\left(y_{n+1}^k - 1\right) & 1+\Delta t\left(\alpha y_{n+1}^k + r\right) \end{pmatrix} \cdot$$
$$\left(\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix}^{k} - \begin{pmatrix} x_n \\ y_n \end{pmatrix} - \Delta t \begin{pmatrix} \alpha y_{n+1}^k\left(1-x_{n+1}^k\right) - r x_{n+1}^k \\ \beta x_{n+1}^k\left(1-y_{n+1}^k\right) - \mu y_{n+1}^k \end{pmatrix} \right), \tag{6}$$

moreover

$$\det = 1 + \Delta t\left(\beta x_{n+1}^k + \alpha y_{n+1}^k + r + \mu\right) + \Delta t^2\left(\beta x_{n+1}^k(\alpha + r) + \alpha y_{n+1}^k(\beta+\mu) + r\mu - \alpha\beta\right). \tag{7}$$

This means that for x_{n+1}^{k+1} we get

$$\begin{aligned} &x_{n+1}^{k+1} = x_{n+1}^k - \\ &\frac{1}{\det}\left(1+\Delta t\left(\beta x_{n+1}^k + \mu\right)\right)\left(x_{n+1}^k - x_n - \Delta t\left(\alpha y_{n+1}^k\left(1-x_{n+1}^k\right) - r x_{n+1}^k\right)\right) + \\ &\frac{1}{\det}\Delta t\alpha\left(x_{n+1}^k - 1\right)\left(y_{n+1}^k - y_n - \Delta t\left(\beta x_{n+1}^k\left(1-y_{n+1}^k\right) - \mu y_{n+1}^k\right)\right). \end{aligned} \tag{8}$$

It is reasonable to start the iteration with $x_{n+1}^0 = x_n$, $y_{n+1}^0 = y_n$. We set $x_n = y_n = \alpha = \beta = \epsilon$, $\mu = r = \epsilon^2$ and $\Delta t = \frac{1}{\epsilon^2}$ where ε is a small positive number. We are interested in the positivity of x_{n+1}^1 namely, what can we expect after one step of Newton iteration? We can calculate $\det = 8 + \frac{2}{\epsilon} - \frac{1}{\epsilon^2} = \frac{8\epsilon^2+2\epsilon-1}{\epsilon^2}$, which is negative for small enough ϵ. Consequently,

$$x_{n+1}^1 = \epsilon + \frac{\epsilon(1-4\epsilon^2)}{8\epsilon^2+2\epsilon-1}, \tag{9}$$

which is negative if det is negative and it is easy to see that, $\varepsilon < \frac{1}{4}$ will guarantee it.

As a consequence we reached a negative x_{n+1}^1 and we lost the DDPP. We know that the implicit Euler method would produce a discrete model which possesses this property unconditionally, which means that only the Newton iteration is responsible for this failure. This phenomena is visualized by the following example, see Fig. 1 for which the parameters are given in Table 1.

Further we suggest two possible techniques which are able avoid the problem due to implicitness and at the same time the restriction for the time-step is less severe compared with the EEM (Tables 2 and 3).

The first is the application of non-local discretization methods leading to semi-implicit schemes. To this aim, we start with (4) and we make a small modification: in the first equation at the right hand side y_{n+1} is substituted by y_n.

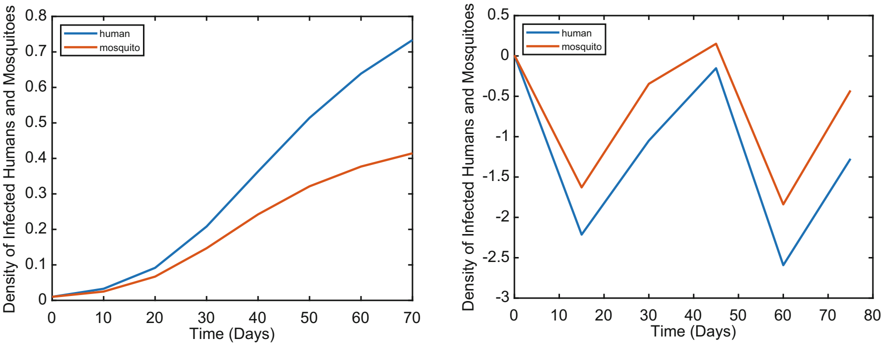


Fig. 1 The impact of the Newton iteration on IEM. (**a**) visualizes that the DDPP is valid with the step size $\Delta t = 10$, while (**b**) shows that the DDPP is not valid with the step size $\Delta t = 15$

Table 1 Parameters and initial values

$x(0) = 0.01$	$y(0) = 0.01$	$r = 0.002$	$\mu = 0.2$	$\alpha = 0.096$	$\beta = 0.2$

Table 2 $\Delta t = 10$, IEM with Newton iteration

–	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
det	1.1702	1.2338	1.4213	1.8649	2.5827	3.3917	4.1254
x^1_{n+1}	0.0327	0.0917	0.2081	0.3631	0.5142	0.6388	0.7336
y^1_{n+1}	0.0248	0.0670	0.1469	0.2423	0.3214	0.3771	0.4144

Table 3 $\Delta t = 15$, IEM with Newton iteration

–	$n=0$	$n=1$	$n=2$	$n=3$
det	−0.1543	35.3487	2.0417	−0.1380
x^1_{n+1}	−2.2137	−1.0495	−0.1519	−2.5924
y^1_{n+1}	−1.6289	−0.3453	0.1514	−1.8389

$$\begin{cases} \dfrac{x_{n+1} - x_n}{\Delta t} = \alpha y_n(1 - x_{n+1}) - r x_{n+1} \\ \dfrac{y_{n+1} - y_n}{\Delta t} = \beta x_{n+1}(1 - y_{n+1}) - \mu y_{n+1}. \end{cases} \tag{10}$$

From the qualitative point of view, it remains implicit, however we can express x_{n+1} from the first equation and then we can solve the second, too thus, it works as an explicit method.

Theorem *The discrete model* (10) *possesses the DDPP for any step size* Δt.

Proof From the first equation we obtain

$$x_{n+1} = \frac{x_n + \Delta t \alpha y_n}{1 + \Delta t \alpha y_n + \Delta t r} \tag{11}$$

which denominator is positive, and the right hand side $\in [0, 1]$ provided $x_n, y_n \in [0, 1]$. The calculation for y_{n+1} is similar, where we can exploit that we already know that $x_{n+1} \in [0, 1]$. □

The second possibility we show is the application of step-size functions instead of the conventional step size Δt, see [1, 3] for more details, which has the form

$$\begin{cases} \dfrac{x_{n+1} - x_n}{\Phi(\Delta t)} = \alpha y_n(1 - x_n) - r x_n \\ \dfrac{y_{n+1} - y_n}{\Phi(\Delta t)} = \beta x_n(1 - y_n) - \mu y_n \end{cases} \tag{12}$$

if we combine it with the EEM. The idea is to find a suitable function Φ for which the method remains a first order method (as EEM), but its qualitative (and stability) properties become better.

Formally, we can state exactly the same: the DDPP is valid for the discrete model (12), iff the step size $\Phi(\Delta t) \in (0, h^*]$, where h^* is defined at (3). The question is that did we gain something or not?

The answer is yes, a suitable choice is

$$\Phi_C(\Delta t) = \frac{1 - e^{-C\Delta t}}{C}, \tag{13}$$

with $C > 0$. With this choice the method remains first order and at the same time $\Phi_C(\Delta t)$ can be sufficiently small for any arbitrary large step size Δt since this function is monotonically decreasing in C and

$$\lim_{C\to 0+} \Phi_C(\Delta t) = \Delta t \quad \text{and} \quad \lim_{C\to +\infty} \Phi_C(\Delta t) = 0. \tag{14}$$

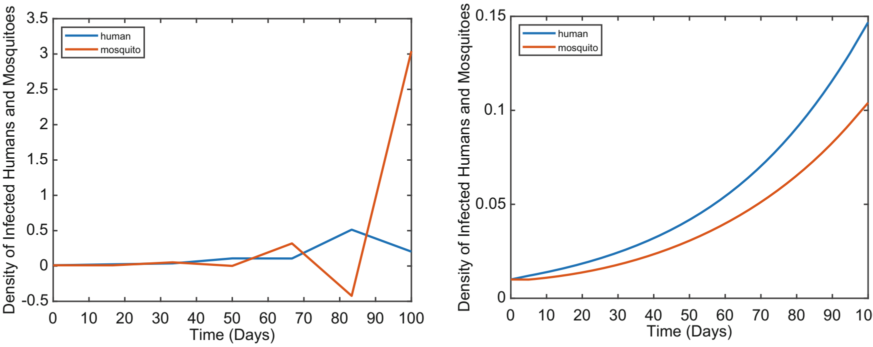
Hence, if

1. $\Delta t < h^*$, then $\Phi_C(\Delta t) \in (0, h^*]$ for all $C > 0$.
2. $\Delta t \geq h^*$, there exists C_0 such that $\Phi_C(\Delta t) \in (0, h^*]$ for $C \geq C_0$.

Here C_0 is the solution of

$$e^{-C_0\Delta t} + C_0 h^* - 1 = 0, \tag{15}$$

implying the sufficient condition $C_0 \geq \frac{1}{h^*}$ and the exact value for C_0 lies on the principal branch of the Lambert W function. This means that for an arbitrarily large Δt we can find a suitable C for which (12) will satisfy the DDPP. See Fig. 2.



(a) Standard EEM, $\Delta t = 15$ (b) EEM, $\Phi_{0.4}(15)$

Fig. 2 The explicit Euler method applied to the Ross-Macdonald model with $\Delta t = 15$. In this example, the step size tolerance $\Delta t^* = 5$ and the parameters are given in Table 1. (**a**) shows that the DDPP is not satisfied, as expected; (**b**) shows that with the application of the step size function $\Phi_{0.4}(15)$ the DDPP is valid

References

1. Anguelov, R., Lubuma, J. M.-S.: Contributions to the Mathematics of the Nonstandard Finite Difference Method and Applications. Numerical Methods for Partial Differential Equations 17, 518–543 (2001).
2. Faragó, I., Mincsovics, M.E., Mosleh, R.: Reliable Numerical Modelling of Malaria Propagation. Application of Mathematics, Springer, 63, 259–271 (2018).
3. Mickens, R.E.: Advances in the Applications of Nonstandard Finite Difference Schemes. World Scientific Publishing Co. Pte. Ltd, (2005).

Dynamics of a Delayed Kaldor-Kalecki Model of Mutually Linked Economies

Juancho A. Collera and Rainier Ric B. de la Cruz

Abstract The Kaldor-Kalecki business cycle model unifies the concept of nonlinearity of the investment function of Kaldor and the concept of investment lag of Kalecki. Recently, the influence of a global economy on a local economy was investigated using delayed Kaldor-Kalecki models that are coupled unidirectionally. In this work, we proposed and analyzed a business cycle model of two mutually linked economies described using bidirectionally coupled Kaldor-Kalecki models. For the case of comparable economies, we obtain an equivariant system. This symmetry property is then used to classify the Hopf bifurcations and consequently determine the different types of oscillatory patterns that can occur in the proposed business cycle model. Our result is further expanded using numerical continuation which detected the occurrence of various limit-cycle bifurcations including period-doubling and torus bifurcations which give rise to different kinds of oscillations. This diversity in oscillatory behavior is crucial to better depict economic cycles.

1 Introduction

Since the inception of Kaldor's trade cycle model [7], a number of researchers have modified this model to account for Kalecki's idea of gestation period in investments. See, for example, the seminal work of Szydłowski and Krawiec in [10] and the delayed Kaldor-Kalecki model introduced by Kaddar and Talibi Alaoui in [6] which incorporates a time delay in both the capital stock and the gross domestic product in

J. A. Collera (✉)
Department of Mathematics and Computer Science, University of the Philippines Baguio, Baguio, Philippines
e-mail: jacollera@up.edu.ph

R. R. B. de la Cruz
Department of Economics and Political Science, University of the Philippines Baguio, Baguio, Philippines
e-mail: rbdelacruz1@up.edu.ph

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_12

the capital accumulation equation. Recently, Jackowska-Zduniak et al. investigated the influence of a global economy on a local economy using two delayed Kaldor-Kalecki models that are coupled unidirectionally [5]. They showed that, as a result of the unidirectional coupling, the powerful global system can restore the weaker market's oscillatory character and oscillations in both systems can have the same period. The case where economies are mutually linked and are comparable in size was not considered by previous works, and hence it is a gap that needs to be addressed.

In this paper, we propose and analyze a business cycle model of two mutually linked economies described using bidirectionally coupled Kaldor-Kalecki models. For the case of comparable economies, we obtain an equivariant system. This symmetry property is then used to classify the different types of oscillatory patterns that can occur in the system. This result is further expanded using numerical continuation which detects various limit-cycle bifurcations showing that the proposed model exhibits richer oscillatory behavior which could better represent economic cycles. The paper is organized as follows. In Sect. 2 we revisit previous works on a delayed Kaldor-Kalecki model, and then in Sect. 3 we introduce and analyze our proposed model of mutually linked economies. We end with conclusions and future directions of the paper in Sect. 4.

2 Delayed Kaldor-Kalecki Model of Business Cycle

In this section, we revisit the delayed Kaldor-Kalecki model studied in [6]. This model is a special case of our proposed model and will be a basis of comparison for our results in Sect. 3. The model first introduced by Kaddar and Talibi Alaoui in [6] incorporates the time delay in both the capital stock K and the gross domestic product Y in the capital accumulation equation and is given as follows

$$\begin{cases} \dot{Y}(t) = \alpha\big[\, I(Y(t), K(t)) - S(Y(t), K(t))\big], \\ \dot{K}(t) = I(Y(t-\tau), K(t-\tau)) - \delta K(t). \end{cases} \tag{1}$$

The savings function $S(Y, K) = \sigma Y$ with $\sigma \in (0, 1)$, while the investment function $I(Y, K) = F(Y) - \beta K$, where $\beta > 0$ and F is a sigmoid function. This shape of the graph of F follows the assumption in [7] that $\partial I/\partial Y$ will be small both for low and for high levels of Y. The positive parameters α and δ correspond respectively to the adjustment coefficient in the goods market and the depreciation rate of the capital stock, while the time delay parameter $\tau > 0$ reflects the delay in investment processes. A unique positive equilibrium solution $E^+ := (Y^*, K^*)$ of system (1) was shown to exist provided $F(0) > 0$, and for all $Y \in \mathbb{R}$, we have $F'(Y) - \sigma < \sigma\beta/\delta$ and $|F(Y)| \leq L$ for some constant $L > 0$. Moreover, if $|F'(Y^*) - \sigma| < \sigma\beta/\delta$ and $F'(Y^*) - \sigma < (\beta + \delta)/\alpha$, then there exists a critical time delay value $\tau_0 > 0$ such that E^+ is locally asymptotically stable (LAS) when $\tau \in [0, \tau_0)$ and is unstable when $\tau > \tau_0$. At $\tau = \tau_0$, system (1) undergoes a Hopf bifurcation (HB) at E^+, and a family of limit cycles (LCs) bifurcates.

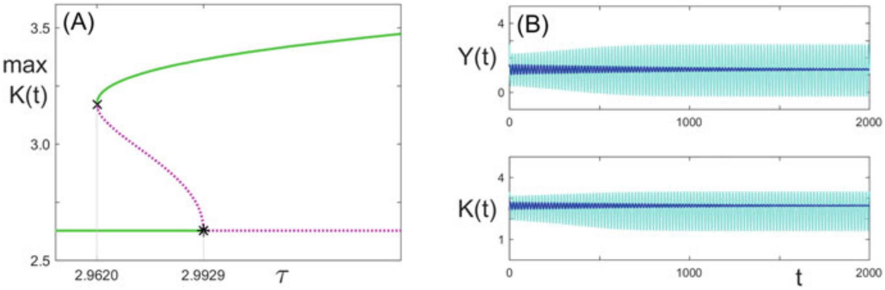


Fig. 1 (**a**) Branch of limit cycles (LCs) emerging from the Hopf bifurcation at $\tau = \tau_0 \approx 2.9929$. (**b**) Time series plots showing the coexistence of stable equilibrium and LC solutions at $\tau = 2.9750$

In order to further reveal the dynamics of system (1), we now carry out numerical continuation using $F(Y) = e^Y/(1 + e^Y)$ and $(\alpha, \beta, \sigma, \delta) = (3, 0.2, 0.2, 0.1)$ as in [6]. For purposes of comparison, we will use the same function and parameter values to our proposed model in Sect. 3. This choice of the sigmoid function and parameter values was shown to satisfy the conditions for the existence of a unique positive equilibrium $E^+ \approx (1.3135, 2.6270)$ and the conditions for HB to occur in system (1) at $\tau = \tau_0 \approx 2.9929$ [6]. Figure 1a shows the branches of solutions obtained using numerical continuation in *DDE-Biftool* [4, 8] varying the time delay parameter τ. The stable and unstable parts of these branches are shown in solid green and dotted magenta lines, respectively. The switch in stability along the branch of equilibria (horizontal line) occurred at a HB marked with ($*$), while the switch in the stability along the branch of LCs (curve) occurred at a saddle-node (SN) bifurcation of LCs marked with ($\times$). This numerical simulation complements the results in [6]. It also reveals the possibility of a coexistence of a stable equilibrium solution and a stable limit-cycle solution as shown in the time series plots in Fig.1b obtained using two different sets of history functions when $\tau = 2.9750$. This multitype bistability occurs for values of τ between 2.9620 and 2.9929 where the SN bifurcation of LCs and the HB occurred approximately.

3 Business Cycle Model of Mutually Linked Economies

Our proposed business cycle model of mutually linked economies is the following

$$\begin{cases} \dot{Y}_1(t) = \alpha\big[F(Y_1(t)) - \beta K_1(t) - \sigma Y_1(t)\big], \\ \dot{K}_1(t) = F(Y_1(t-\tau)) - \beta K_1(t-\tau) - \delta K_1(t) + \eta_1\big[Y_2(t) - Y_1(t)\big], \\ \dot{Y}_2(t) = \alpha\big[F(Y_2(t)) - \beta K_2(t) - \sigma Y_2(t)\big], \\ \dot{K}_2(t) = F(Y_2(t-\tau)) - \beta K_2(t-\tau) - \delta K_2(t) + \eta_2\big[Y_1(t) - Y_2(t)\big]. \end{cases} \tag{2}$$

This system is composed of two coupled Kaldor-Kalecki models similar to system (1). Hence, the description of the state variables, parameters, and functions are similar to the ones given in system (1) with the addition of the coupling parameters $\eta_1, \eta_2 > 0$. Here, we assume that the two economies are comparable and the coupling is bidirectional with identical coupling coefficients, i.e. $\eta_1 = \eta_2$. We denote this common coupling parameter simply by η. This symmetric case is significant because it organizes the dynamics of systems with almost identical Kaldor-Kalecki models, i.e. systems similar to system (2) that have different but almost identical corresponding parameters in each Kaldor-Kalecki model. The symmetry assumptions on system (2) make it an *equivariant system under some symmetry group* Γ. That is, if we write system (2) in the form $\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X}_t)$ where $\mathbf{X}(t) = [Y_1(t), K_1(t), Y_2(t), K_2(t)]^\top$ and $\mathbf{X}_t \in C([-\tau, 0], \mathbb{R}^4)$, the space of continuous functions mapping $[-\tau, 0]$ into $\mathbb{R}^4$, with $\mathbf{X}_t(\theta) = \mathbf{X}(t+\theta)$ for $\theta \in [-\tau, 0]$, then $\gamma \cdot \mathbf{F}(\mathbf{X}_t) = \mathbf{F}(\gamma \cdot \mathbf{X}_t)$ for all $\gamma \in \Gamma$. It is straightforward to show that system (2) has symmetry group $\Gamma \cong \mathbb{Z}_2 = \langle \gamma \rangle$ where its action on the state variables is given by $\gamma \cdot (Y_1, K_1, Y_2, K_2) = (Y_2, K_2, Y_1, K_1)$. The solutions fixed by Γ must have $Y_1(t) = Y_2(t)$ and $K_1(t) = K_2(t)$. In particular, one can obtain the so-called *fully symmetric equilibrium* $E^* := (Y^*, K^*, Y^*, K^*)$ where the components Y^* and K^* are exactly the same components as that of E^+ from Sect. 2. This is because if $Y_1(t) = Y_2(t)$ and $K_1(t) = K_2(t)$, then the coupling terms in system (2) vanish and we end up having two copies of system (1).

We now examine the local stability of the equilibrium E^* and its bifurcations. Using symmetry techniques, see e.g. references [2, 3, 9], we classify the types of periodic solutions that can arise in our business cycle model of linked economies. The characteristic equation corresponding to the linearized system about E^* is

$$\det\bigl(\Delta(\lambda)\bigr) = 0 \qquad \text{with} \qquad \Delta(\lambda) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \tag{3}$$

and where $\mathbf{A}$ and $\mathbf{B}$ are 2×2 matrices. The action of Γ on $\mathbb{R}^4$ yields the isotypic decomposition $\mathbb{R}^4 = \mathbb{T}^2 \oplus \mathbb{A}^2$ which allows block diagonalizing $\Delta(\lambda)$ into the form $\mathrm{diag}(\mathbf{A}+\mathbf{B}, \mathbf{A}-\mathbf{B})$. Hence, we have that $\det(\Delta(\lambda)) = \det(\mathbf{A}+\mathbf{B}) \cdot \det(\mathbf{A}-\mathbf{B})$ and Eq. (3) splits into the following equations

$$\det(\mathbf{A}+\mathbf{B}) = \lambda^2 + \alpha(\sigma - \rho)(\lambda + \delta) + \delta\lambda + (\lambda + \alpha\sigma)\beta e^{-\lambda\tau} = 0, \tag{4}$$

$$\det(\mathbf{A}-\mathbf{B}) = \lambda^2 + \alpha(\sigma - \rho)(\lambda + \delta) + \delta\lambda + (\lambda + \alpha\sigma)\beta e^{-\lambda\tau} - 2\alpha\beta\eta = 0, \tag{5}$$

where $\rho = F'(Y^*)$. Moreover, the critical roots of (4) give rise to regular bifurcations while the critical roots of (5) give rise to symmetry-breaking bifurcations. Since the form of (4) and (5) is the same as the characteristic equation obtained in [6], we get analogous results. That is, if the equilibrium E^* is LAS when $\tau = 0$, then a switch in the stability of E^* can occur at some critical time delay $\tau_c = \min\{\tau_0^+, \tau_0^-\}$ where τ_0^+ and τ_0^-, respectively, are the least positive time delay values

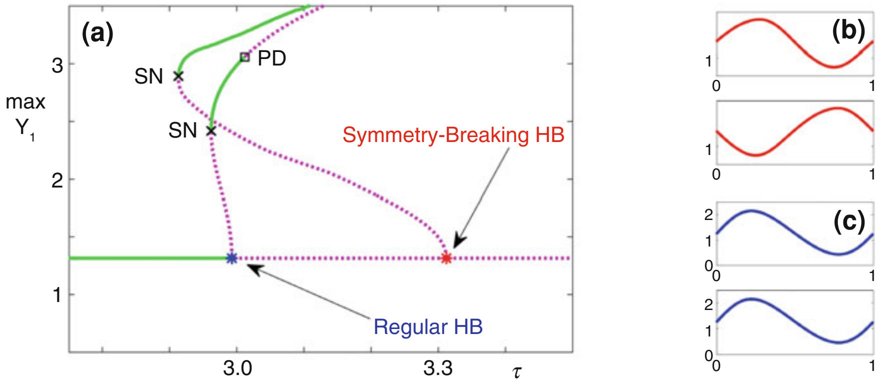


Fig. 2 (**a**) HBs along the branch of fully symmetric equilibria (horizontal line) and the branches of LCs emerging from these HBs. Patterns of oscillation of LC solutions emerging from (**b**) the symmetry-breaking HB and (**c**) the regular HB

where Eqs. (4) and (5) have simple purely imaginary roots with $\frac{d}{d\tau}\mathrm{Re}\lambda(\tau)\big|_{\tau=\tau_0^{\pm}} > 0$. In this case, system (2) undergo a regular HB at the equilibrium E^* when $\tau = \tau_0^+$ and a symmetry-breaking HB at E^* when $\tau = \tau_0^-$.

The following example illustrates the classification of periodic solutions to our business cycle model of mutually linked economies. For system (2) with $\eta = 0.10485$ and the same function F and parameter set as in Sect. 2, we obtain $\tau_0^+ \approx 2.9929$ and $\tau_0^- \approx 3.3125$. Observe that the value τ_0^+ is the same as the value τ_0 in Sect. 2. This is because (4) is independent of η and is exactly the same as the characteristic equation in [6]. In contrast, Eq. (5) depends on η. Figure 2a shows the regular and symmetry-breaking HBs along the branch of fully symmetric equilibria (horizontal line) which occurred at $\tau = \tau_0^+$ and $\tau = \tau_0^-$, respectively. The patterns of oscillation of LC solutions emerging from these HBs are of two types. LCs from the symmetry-breaking HB as represented in Fig. 2b show economies oscillating in an asynchronous manner, while LCs from the regular HB as represented in Fig.2c show synchrony between oscillating economies. Numerical continuation from both HBs yield branches of LCs having stable parts. That is, the proposed model supports both synchronous and asynchronous oscillations. Figure 2a shows the branch of LCs that emerged from the regular HB. Compared to the branch of LCs shown in Fig. 1a for system (1), the LCs emanating from the regular HB in Fig. 2a switch stability for the second time at a period-doubling (PD) bifurcation marked with (□).

Figure 3a shows the branch of LCs emerging from the PD bifurcation at $\tau = \tau_{PD} \approx 3.0119$. A stability switch along this period-2 LC branch occurs at a torus bifurcation (TB) marked with (△) at $\tau = \tau_{TB} \approx 3.0394$. Observe that for $\tau \in (\tau_{PD}, \tau_{TB})$, a stable period-1 and a stable period-2 LCs coexist. Figure 3b and c shows the profile plot of the coexisting LCs for $\tau = 3.0250$. Figure 4a and b shows the corresponding Floquet multipliers before and after the PD bifurcation at $\tau = \tau_{PD}$. This establishes the existence of the PD bifurcation and the switch towards

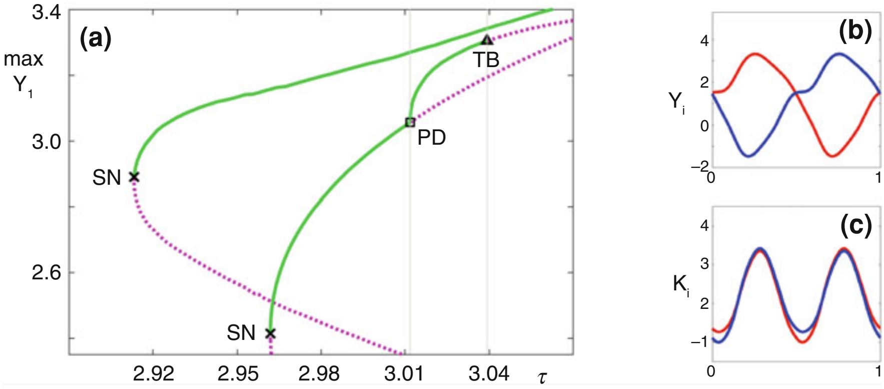


Fig. 3 (**a**) PD bifurcation along the branch of LC solutions that emerged from the regular HB. Profile plot of coexisting (**b**) period-1 and (**c**) period-2 stable LC solutions at $\tau = 3.0250$

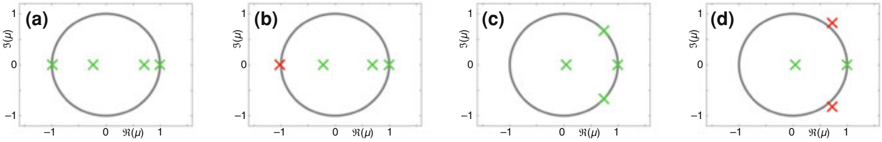


Fig. 4 Floquet multipliers (**a**) before and (**b**) after the PD bifurcation at $\tau = \tau_{PD}$ along the LC branch emanating from the regular HB, and Floquet multipliers (**c**) before and (**d**) after the TB at $\tau = \tau_{TB}$ where LCs on the period-2 branch switch stability

instability at $\tau = \tau_{PD}$ along the LC branch from the regular HB. Similarly, the existence of the TB at $\tau = \tau_{TB}$ and the stability switch along the period-2 LC branch are corroborated in Fig. 4c and d showing the corresponding Floquet multipliers before and after the TB at $\tau = \tau_{TB}$.

When the common coupling parameter η in system (2) is zero, the two Kaldor-Kalecki models decouple and the dynamics of the individual economy are identical to the one discussed in Sect. 2. This dynamical behavior extends to the cases where the value of η is sufficiently small. However, for larger values of η, a variety of dynamics can be obtained. For brevity, we only presented the dynamics for the case where $\eta = 0.10485$ because this particular case illustrates the occurrence of various LC bifurcations including PD and torus bifurcations. These LC bifurcations consequently render some form of irregularity compared to the usual approach of using smooth limit cycles to describe economic fluctuations. Furthermore, this result supports the idea in [1] that endogenous cycles buffeted by exogenous disturbances, in our case the effects of mutual coupling, may result to irregular fluctuations. In other words, the obtained diversity in oscillatory behavior better depicts fluctuations in economies.

4 Conclusions and Future Directions

We introduced a business cycle model of two mutually linked economies using bidirectionally coupled Kaldor-Kalecki models. For comparable economies, the resulting system is equivariant and this symmetry property played an important role in our analysis. In particular, a classification of Hopf bifurcations into regular and symmetry-breaking was provided. Numerical continuation from both the regular and the symmetry-breaking HBs yield branches of LCs having stable parts. That is, the proposed model supports both synchronous and asynchronous oscillations. Moreover, we showed that various types of limit-cycle bifurcations can occur in the proposed model including period-doubling and torus bifurcations which give rise to different kinds of oscillations. This diversity in oscillatory behavior is crucial to better depict economic cycles. For example, the occurrence of a torus bifurcation yields quasiperiodic oscillations which could mimic fluctuations in the economies. To better understand the effects of coupling between economies, a more general form of the proposed model with different coupling parameters is currently being considered and is the subject of an on-going research.

Acknowledgments The authors would like to thank the referees for their valuable review that greatly improved the quality of this paper. JAC also acknowledges the support of the University of the Philippines Baguio through RLCs during the Second Semester of A.Y. 2021–2022.

References

1. Beaudry, P., Galizia, D., Portier, F.: Reviving the limit cycle view of macroeconomic fluctuations. NBER Working Papers 21241 (2015)
2. Collera, J.A.: Symmetry-breaking bifurcations in two mutually delay-coupled lasers. Phil. Sci. Tech. **8**, 17–21 (2015)
3. Collera J.A.: Queues with choice from a symmetry perspective. In: Faragó I., et al. (eds) Progress in Industrial Mathematics at ECMI 2018, pp. 537–542. Springer, Cham (2019)
4. Collera J.A.: Numerical continuation and bifurcation analysis in a harvested predator-prey model with time delay using DDE-Biftool. In: Mohd M.H., et al. (eds) Dynamical Systems, Bifurcation Analysis and Applications, pp. 225–241. Springer, Singapore (2019)
5. Jackowska-Zduniak, B., Grzybowska, U., Orłowski, A.: Numerical analysis of two coupled Kaldor-Kalecki models with delay. Acta Phys. Pol. A **127**, A70–A74 (2015)
6. Kaddar, A., Talibi Alaoui, H.: Hopf bifurcation analysis in a delayed Kaldor-Kalecki model of business cycle. Nonlinear Anal-Model **13**, 439–449 (2008)
7. Kaldor, N.: A model of the trade cycle. Econ. J. **50**, 78–92 (1940)
8. Luzyanina, T., Sieber, J., Engelborghs, K., Samaey, G., Roose, D.: Numerical bifurcation analysis of mathematical models with time delays with the package DDE-BIFTOOL. Mat. Biolog. Bioinform. **12**, 496–520 (2017)
9. Palacios, A.: Synchronization in asymmetrically coupled networks with homogeneous oscillators. Phys. Rev. E **103**, 022206 (2021)
10. Szydłowski, M., Krawiec, A.: The Kaldor–Kalecki model of business cycle as a two-dimensional dynamical system. J. Nonlinear Math. Phys. **8**, 266–271 (2001)

Dynamic Iterations for Nonlinear Systems Applied in Population Dynamics

Barbara Zubik-Kowal

Abstract We investigate the application of dynamic iterations to nonlinear systems of differential equations. Such an application allows to use implicit time integration methods without solving nonlinear algebraic equations at each time step. Another advantage of the application of dynamic iterations is that the resulting numerical schemes can be solved in parallel computing environments. We conclude that the sequence of how the dynamic iterations are applied is significant and influences their rate of convergence to the solution of the given system of nonlinear differential equations. This conclusion is illustrated by numerical experiments involving Volterra equations for predator-prey interactions. We also conclude that the proposed numerical scheme is faster than the variable order method.

1 Introduction

Dynamic iterations have been broadly investigated as numerical methods applied to solve differential systems on parallel computers and are often called waveform relaxation techniques. These techniques have been introduced by Lelarasmee et al. [4] and investigated by many authors for different kinds of differential equations, see, for example, [3] and [5] for systems of ordinary differential equations, [1] and [2] for systems of delay differential equations and [6] and [7] for general functional differential equations. However, these techniques have been mainly investigated in the context of parallel computations and not much attention has been given to answer the question of whether or not permutations of the equations in a given system influence the convergence of the applied dynamic iterations. Recent investigations [8] in this direction for linear systems of differential equations show that appropriately chosen permutations, in light of the values of the model parameters, present a way to speed up the convergence of dynamic iterations.

B. Zubik-Kowal (✉)
Department of Mathematics, Boise State University, Boise, ID, USA
e-mail: bzubik@boisestate.edu

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_13

The goal of the current paper is to address this question for nonlinear differential equations.

In this paper, we investigate dynamic iterations for systems written in the form

$$\begin{cases} \dfrac{d}{dt}x = ax + xf_1(y) + g_1(t) \\ \dfrac{d}{dt}y = \tilde{a}y + yf_2(x) + g_2(t) \end{cases} \tag{1}$$

where a, $\tilde{a}$ are real parameters and f_i, g_i, $i = 1, 2$, are given real functions. The system (1) is supplemented by the initial conditions

$$x(0) = \xi_0, \quad y(0) = \eta_0.$$

For an arbitrary continuous function $y^{(0)}(t)$, we consider the sequences $\{x^{(k)}(t)\}_{k=1}^{\infty}$, $\{y^{(k)}(t)\}_{k=0}^{\infty}$ defined by

$$\begin{cases} \dfrac{d}{dt}x^{(k+1)} = ax^{(k+1)} + x^{(k+1)}f_1(y^{(k)}) + g_1(t), \\ \dfrac{d}{dt}y^{(k+1)} = \tilde{a}y^{(k+1)} + y^{(k+1)}f_2(x^{(k+1)}) + g_2(t), \end{cases} \tag{2}$$

where $k = 0, 1, 2, \ldots$ and

$$x^{(k+1)}(0) = \xi_0, \quad y^{(k+1)}(0) = \eta_0.$$

The numerical scheme (2) is called Gauss-Seidel waveform relaxation. The advantage of (2) over (1) is that the application of implicit methods, for example BDF methods, for integration of (2) in time t does not require solving nonlinear algebraic equations at each time step. Note that the application of implicit time integration methods to the nonlinear differential system (1) leads to a system of nonlinear algebraic equations that require an additional process to solve them at each time step (more time steps mean that more nonlinear algebraic systems need to be solved). Such an additional process would not be needed if system (2) would be applied.

The paper is organized as follows. In Sect. 2, we analyze the convergence of the sequence $\{(x^{(k)}(t), y^{(k)}(t))\}_{k=0}^{\infty}$ to the exact solution $(x(t), y(t))$ as $k \to \infty$. Then, in Sect. 3, we present results of numerical experiments involving nonlinear systems applied in population dynamics. Finally, we finish with concluding remarks in Sect. 4.

2 Error Analysis

In this section, we analyze the errors

$$e_x^{(k)}(t) = x^{(k)}(t) - x(t), \quad e_y^{(k)}(t) = y^{(k)}(t) - y(t),$$

as $k \to \infty$. We assume that the unknown solutions x, y are bounded and the given functions f_1, f_2 are Lipschitz continuous. In what follows, we use the following notation. Let L_1, L_2 be Lipschitz constants for f_1 and f_2, respectively; that is,

$$\begin{aligned} |f_1(y) - f_1(\tilde{y})| &\leq L_1|y - \tilde{y}|, \quad \text{for all } y, \tilde{y} \in \mathbb{R}, \\ |f_2(x) - f_2(\tilde{x})| &\leq L_2|x - \tilde{x}|, \quad \text{for all } x, \tilde{x} \in \mathbb{R} \end{aligned}$$

and X, Y, R, E_0 be positive constants such that

$$\begin{aligned} &|x(t)| \leq X, \quad |y(t)| \leq Y, \quad \text{for all } 0 \leq t, \\ &f_1(y) + a \leq R, \quad f_2(x) + \tilde{a} \leq R, \quad \text{on bounded sets}, \\ &|e_y^{(0)}(t)| \leq E_0, \quad \text{for all } 0 \leq t. \end{aligned}$$

The following theorem provides error bounds for scheme (2).

Theorem 1 *Let* $k = 0, 1, 2, \ldots$ *and* $t \geq 0$. *Then,*

$$|e_x^{(k+1)}(t)| \leq E_0 X L_1 \big(XYL_1L_2\big)^k \frac{e^{Rt}}{R^{2k+1}} \sum_{j=2k+1}^{\infty} \frac{(-1)^{j+1}(Rt)^j}{j!}, \tag{3}$$

$$|e_y^{(k+1)}(t)| \leq E_0 \big(XYL_1L_2\big)^{k+1} \frac{e^{Rt}}{R^{2k+2}} \sum_{j=2k+2}^{\infty} \frac{(-1)^{j}(Rt)^j}{j!}. \tag{4}$$

Proof Note that $e_x^{(k)}(0) = 0$ and $e_y^{(k)}(0) = 0$, for all $k = 0, 1, 2, \ldots$. Then, from (1) and (2), we get

$$|e_x^{(k+1)}(t)| \leq XL_1 \int_0^t |e_y^{(k)}(\tau)| e^{R(t-\tau)} d\tau, \tag{5}$$

$$|e_y^{(k+1)}(t)| \leq YL_2 \int_0^t |e_x^{(k+1)}(\tau)| e^{R(t-\tau)} d\tau, \tag{6}$$

for $k = 0, 1, 2, \ldots$ and $t \geq 0$. From (5), we get

$$\begin{aligned} |e_x^{(1)}(t)| &\leq XL_1E_0 \int_0^t e^{R(t-\tau)} d\tau = \frac{XL_1E_0}{R}\Big(e^{Rt} - 1\Big) = \frac{XL_1E_0}{R} e^{Rt}\Big(1 - \sum_{j=0}^{\infty} \frac{(-Rt)^j}{j!}\Big) \\ &= \frac{XL_1E_0}{R} e^{Rt} \sum_{j=1}^{\infty} \frac{(-1)^{j+1}(Rt)^j}{j!}, \end{aligned}$$

which shows (3) for $k = 0$. We now use (6) and get

$$
\begin{aligned}
|e_y^{(1)}(t)| &\le \frac{XYL_1L_2E_0}{R}e^{Rt}\int_0^t \sum_{j=1}^{\infty}\frac{(-1)^{j+1}(R\tau)^j}{j!}d\tau \\
&= \frac{XYL_1L_2E_0}{R}e^{Rt}\sum_{j=1}^{\infty}\frac{(-1)^{j+1}R^jt^{j+1}}{(j+1)!} = \frac{XYL_1L_2E_0}{R^2}e^{Rt}\sum_{j=2}^{\infty}\frac{(-Rt)^j}{j!},
\end{aligned}
$$

which shows (4) for $k=0$. We now assume (3) and (4) for a certain k. Then, from (5) and (4), we get

$$
\begin{aligned}
|e_x^{(k+2)}(t)| &\le XL_1\int_0^t \big(XYL_1L_2\big)^{k+1}E_0\frac{e^{R\tau}}{R^{2k+2}}\sum_{j=2k+2}^{\infty}\frac{(-R\tau)^j}{j!}e^{R(t-\tau)}d\tau \\
&= XL_1\big(XYL_1L_2\big)^{k+1}E_0\frac{e^{Rt}}{R^{2k+2}}\sum_{j=2k+2}^{\infty}\frac{(-R)^jt^{j+1}}{(j+1)!} \\
&= XL_1\big(XYL_1L_2\big)^{k+1}E_0\frac{e^{Rt}}{R^{2k+3}}\sum_{j=2k+3}^{\infty}\frac{(-1)^{j+1}(Rt)^j}{j!},
\end{aligned}
$$

which, by mathematical induction, shows (3). We now use (6) and (3) and obtain the following result,

$$
\begin{aligned}
|e_y^{(k+2)}(t)| &\le YL_2\int_0^t XL_1\big(XYL_1L_2\big)^{k+1}E_0\frac{e^{R\tau}}{R^{2k+3}}\sum_{j=2k+3}^{\infty}\frac{(-1)^{j+1}(R\tau)^j}{j!}e^{R(t-\tau)}d\tau \\
&= \big(XYL_1L_2\big)^{k+2}E_0\frac{e^{Rt}}{R^{2k+4}}\sum_{j=2k+4}^{\infty}\frac{(-Rt)^j}{j!},
\end{aligned}
$$

which shows (4) and finishes the proof. □

3 Numerical Experiments and Methods Comparison

In this section, we present results of numerical experiments involving dynamic iterations applied to Volterra equations for predator-prey interactions.

The system of interest is of the form

$$
\begin{cases}
\dfrac{d}{dt}x = ax - bxy + g_1(t) \\
\dfrac{d}{dt}y = -cy + dxy + g_2(t)
\end{cases}
\tag{7}
$$

where $a = 2/3$, $b = 20$, $c = 50$, $d = 0.01$, $0 \le t \le 10$. We apply (2) and obtain the following scheme

$$\begin{cases} \dfrac{d}{dt}x^{(k+1)} = ax^{(k+1)} - bx^{(k+1)}y^{(k)} + g_1(t) \\ \dfrac{d}{dt}y^{(k+1)} = -cy^{(k+1)} + dx^{(k+1)}y^{(k+1)} + g_2(t). \end{cases} \tag{8}$$

If we write the equations in system (7) in the opposite order and then apply (2), we obtain a different scheme of the form

$$\begin{cases} \dfrac{d}{dt}y^{(k+1)} = -cy^{(k+1)} + dx^{(k)}y^{(k+1)} + g_2(t) \\ \dfrac{d}{dt}x^{(k+1)} = ax^{(k+1)} - bx^{(k+1)}y^{(k+1)} + g_1(t). \end{cases} \tag{9}$$

Numerical solutions $x^{(k)}(t_n)$ and $y^{(k)}(t_n)$ computed by (8) and (9) are presented in Fig. 1 as functions of t_n for $k = 6$ in the case of (8) and for $k = 4$ in the case of (9).

Although both schemes (8) and (9) originate from the same application of (2), their errors are different and demonstrate different convergence rates. The errors of both schemes are presented in Fig. 2. The upper subplot demonstrates the errors resulting from the application of (8) and the lower subplot demonstrates the errors resulting from the application of (9). Each scheme is integrated by BDF3 with $h = 10^{-4}$.

We now compare the accuracy and CPU time using the solver `ode15s` and scheme (9) integrated by BDF6 with $h = 10^{-2}$. The maximum error

$$\max\left\{\max_n \left|x_1(t_n) - x_{1,n}^{(k)}\right|, \max_n \left|x_2(t_n) - x_{2,n}^{(k)}\right|\right\}$$

is $7.03 \cdot 10^{-13}$ and $1.91 \cdot 10^{-14}$ using the solver `ode15s` and (9), respectively. The errors resulting from the application of both methods are comparable. The CPU

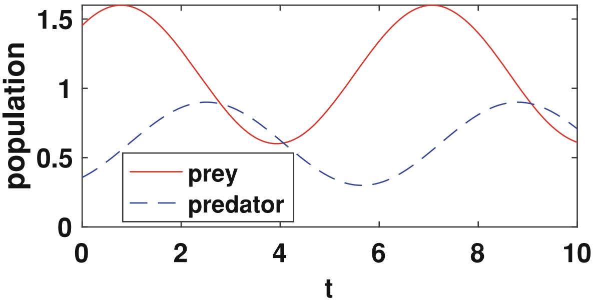


Fig. 1 Numerical solutions of (7): $x(t)$ solid and $y(t)$ dashed

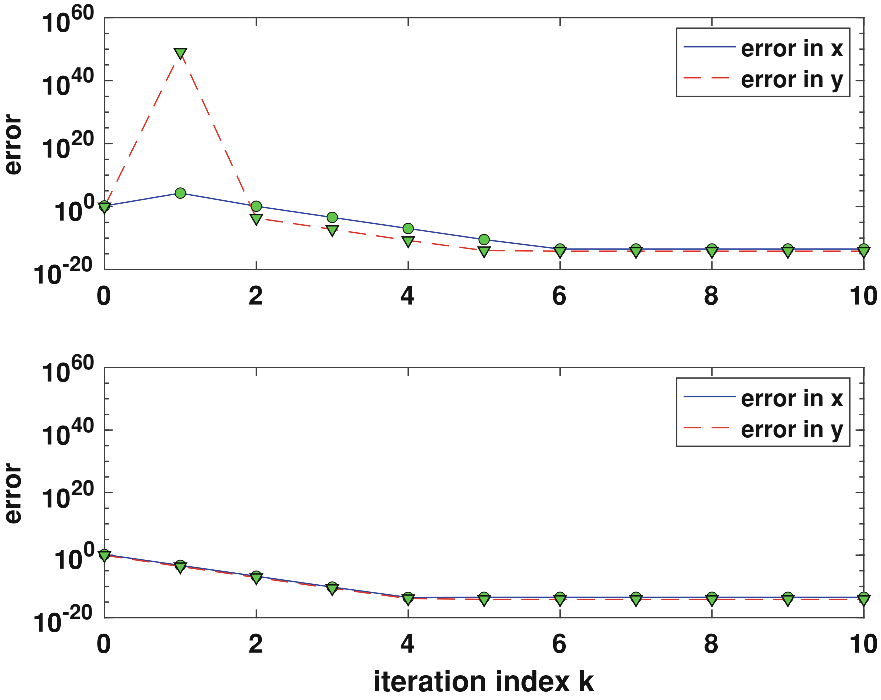


Fig. 2 Methods comparison: errors by (8) (upper subplot) and errors by (9) (lower subplot)

time is 0.21 s when the solver `ode15s` is applied and it is 0.03 s when the numerical scheme (9) is applied, demonstrating that (9) is faster than `ode15s`.

4 Conclusions

In this work, we investigate dynamic iterations for nonlinear systems of differential equations applied in population dynamics. The advantage of dynamic iterations is that they allow to apply implicit time integration methods without the cost of solving nonlinear algebraic equations at each time step. We conclude that the convergence of dynamic iterations is different if we swap the order of the nonlinear differential equations in the given system even though the iterations are applied to the same system. That is, only by swapping the order of the equations, we can increase the rate of the convergence of the iterations. We also conclude that after choosing the optimal permutation of the equations, the proposed numerical scheme based on dynamical iterations is faster than the variable order method.

References

1. Bjorhus, M.: On dynamic iteration for delay differential equations. BIT **43**, 325–336 (1994)
2. Bjorhus, M.: A note on the convergence of discretized dynamic iteration. BIT **35**, 291–296 (1995)
3. Burrage, K.: Parallel and Sequential Methods for Ordinary Differential Equations. Oxford University Press, Oxford (1995)
4. Lelarasmee, E., Ruehli, A., and Sangiovanni-Vincentelli, A.: The waveform relaxation method for time-domain analysis of large scale integrated circuits, IEEE Trans. CAD. **1**, 131–145 (1982)
5. Miekkala, U., and Nevanlinna, O.: Iterative solution of systems of linear differential equations. Acta Numerica, 259–307 (1996)
6. Zubik-Kowal, B., and Vandewalle, S.: Waveform relaxation for functional-differential equations. SIAM J. Sci. Comput. **21**, 207–226 (1999)
7. Zubik-Kowal, B.: Error bounds for spatial discretization and waveform relaxation applied to parabolic functional differential equations. J. Math. Anal. Appl. **293**, 496–510 (2004)
8. Zubik-Kowal, B.: Error analysis and the role of permutation in dynamic iteration schemes. Computational and Analytic Methods in Science and Engineering, 239–256, Birkhäuser/Springer, Cham, (2020)

Delay Differential Equations for Epidemic Models with Temporary Immunity

Roland Pulch

Abstract We consider an epidemic model with populations of susceptible, infectious, and recovered (SIR). A temporary immunity is described by a system of delay differential equation (DDEs) with a single delay. Alternatively, we introduce a system of distributed DDEs, where a probability distribution characterises variations in the time span for the loss of immunity. A numerical method is derived for the distributed DDEs by a Gaussian quadrature. We present results of numerical experiments using a beta distribution.

1 Introduction

Modelling of epidemics often applies systems of ordinary differential equations (ODEs) for the dynamics of populations like susceptible, infectious, recovered, or others, see [2]. A system of delay differential equations (DDEs) with a single delay τ was proposed to incorporate a temporary immunity in [1]. Therein, a recovered individual looses the immunity exactly after the time span τ. However, individuals typically loose their immunity at different times in real life.

Variations or uncertainties are frequently described by probability distributions, see [7, 9]. Thus we model the times for the loss of immunity by a probability density function. This approach generates a system of distributed DDEs. We briefly discuss stationary solutions of this model. The included integral can be discretised by a quadrature rule. This discretisation yields an approximative system of DDEs with a finite number of delays, which can be solved by existing numerical methods as in [4]. We use Gaussian quadrature, see [6], due to an optimal property. Moreover, each traditional probability distribution is associated to a Gaussian quadrature rule.

R. Pulch (✉)
Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany
e-mail: roland.pulch@uni-greifswald.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_14

This article is organised as follows. We specify the models with single delay and distributed delay in Sects. 2 and 3, respectively. The numerical method based on Gaussian quadrature is outlined in Sect. 4. We illustrate results of numerical computations in Sect. 5, where a beta distribution is considered.

2 Model with Single Delay

We examine models with susceptible individuals S, infectious individuals I, and recovered individuals R. In [1], a system of DDEs is used to include temporary immunity. This system reads as

$$\begin{aligned} \dot{S}(t) &= -\alpha I(t)S(t) + \beta I(t-\tau), \\ \dot{I}(t) &= \alpha I(t)S(t) - \beta I(t), \\ \dot{R}(t) &= \beta I(t) - \beta I(t-\tau), \end{aligned} \tag{1}$$

with infection rate α and recovery rate β. The recovered individuals loose their immunity exactly after the time $\tau > 0$. Let $\mathbf{x} = (S, I, R)^\top$. Initial values $\mathbf{x}(t) = \mathbf{x}_0(t)$ for $t \in [-\tau, 0]$ are required with a predetermined function $\mathbf{x}_0 \colon [-\tau, 0] \to [0, \infty)^3$. An SIR model with single delay involving a birth rate as well as a death rate was investigated in [8].

The total population is $N(t) = S(t) + I(t) + R(t)$. It follows that N is constant in time. We normalise $N(t) = 1$ for all t without loss of generality. Thus the population R is neglected in an analysis. Initial values are restricted to the range $[0, 1]^3$ now.

Stationary solutions $\mathbf{x}^* = (S^*, I^*, R^*)^\top$ are characterised by the condition

$$(\beta - \alpha S^*)I^* = 0. \tag{2}$$

There are two families of steady state solutions:

1. disease-free steady state

$$S^* = \sigma, \qquad I^* = 0 \qquad \text{for arbitrary } \sigma \in [0, 1], \tag{3}$$

2. endemic steady state

$$S^* = \tfrac{\beta}{\alpha}, \qquad I^* = \iota \qquad \text{for arbitrary } \iota \in [0, 1]. \tag{4}$$

If a solution of an initial value problem converges to a stationary solution, then the values σ in (3) or ι in (4) depend on the choice of both the initial values $\mathbf{x}_0$ and the parameters α, β, τ.

3 Model with Distributed Delay

In the model (1), the assumption of a single delay represents an idealisation. Generally, the time spans for loosing an immunity vary in real life. We propose a system of distributed DDEs

$$\begin{aligned}\dot{S}(t) &= -\alpha I(t)S(t) + \beta J(t),\\ \dot{I}(t) &= \alpha I(t)S(t) - \beta I(t),\\ \dot{R}(t) &= \beta I(t) - \beta J(t),\end{aligned} \tag{5}$$

including the integral term

$$J(t) = \int_{\tau_{\min}}^{\tau_{\max}} g(s)\, I(t-s)\,\mathrm{d}s, \tag{6}$$

with $0 \leq \tau_{\min} < \tau_{\max} \leq \infty$ and a measurable weight function $g \colon D \to \mathbb{R}$. The interval is either $D = [\tau_{\min}, \tau_{\max}]$ for $\tau_{\max} < \infty$ or $D = [\tau_{\min}, \infty)$ for $\tau_{\max} = \infty$. We assume that g is a probability density function associated to a probability distribution. Hence it holds that $g(s) \geq 0$ for all $s \in D$ as well as

$$\int_{\tau_{\min}}^{\tau_{\max}} g(s)\,\mathrm{d}s = 1. \tag{7}$$

Concerning distributed DDEs, typical choices are a uniform distribution or a gamma distribution, see [5]. The exponential distribution can be seen as a special case of the gamma distribution. Now initial values have to be predetermined by a function $\mathbf{x}_0 \colon [-\tau_{\max}, 0] \to [0, \infty)^3$ if $\tau_{\max} < \infty$ or $\mathbf{x}_0 \colon (-\infty, 0] \to [0, \infty)^3$ if $\tau_{\max} = \infty$.

We also normalise the total population to $N(t) = 1$ for all t. Concerning stationary solutions $\mathbf{x}^* = (S^*, I^*, R^*)^\top$, the property (7) implies $J(t) = I^*$ for all t in the integral (6). Again we obtain the condition (2). It follows that the stationary solutions of the model (1) and the model (5) coincide.

4 Numerical Method

The numerical solution of DDEs with a single delay or a finite number of multiple delays was discussed in [4]. We consider a distributed DDE of the type (5). The integral (6) is approximated by a quadrature formula

$$J(t) \approx \sum_{j=1}^{k} w_j\, I(t - \tau_j), \tag{8}$$

Table 1 Probability distributions and Gaussian quadrature rules

Probability distribution	Support	Quadrature rule
Uniform distribution	$[\tau_{\min}, \tau_{\max}]$	Gauss-Legendre
Beta distribution	$[\tau_{\min}, \tau_{\max}]$	Gauss-Jacobi
Exponential distribution	$[0, \infty)$	Gauss-Laguerre
Gamma distribution	$[0, \infty)$	Generalised Gauss-Laguerre

with pairwise different nodes $\{\tau_1, \tau_2, \ldots, \tau_k\} \subset D$ and weights $\{w_1, w_2, \ldots, w_k\} \subset \mathbb{R}$. Inserting the approximation (8) into the system (5) yields DDEs of same dimension and k discrete delays. Now numerical methods for multiple delays can be used.

The computation work significantly grows with the number of delays. Thus the number k should be kept small. We apply Gaussian quadrature rules, see [6]. This quadrature owns an optimal polynomial exactness for a fixed number k of nodes. If the integrand (without weight function) is a polynomial of degree $\leq 2k - 1$, then the approximation is exact. Table 1 itemises the Gaussian quadrature schemes associated to some traditional probability distributions.

5 Numerical Experiments

In the models (1) and (5), we choose the parameters $\alpha = 0.1$ and $\beta = 0.05$. In the integral (6), we apply a beta distribution, i.e., a probability density of the form

$$g(s) = C\,(\tau_{\max} - s)^{\eta}\,(s - \tau_{\min})^{\nu} \tag{9}$$

for $s \in [\tau_{\min}, \tau_{\max}]$ with exponents $\eta, \nu \geq 0$ and a constant $C > 0$ for standardisation. We always select $\eta = 3$ and $\nu = 1$ in the following, which implies an asymmetric probability distribution. Still $\tau_{\min}, \tau_{\max}$ have to be chosen. The expected value of the random delay reads as

$$\mu = \tau_{\min} + (\tau_{\max} - \tau_{\min})\tfrac{\nu+1}{\eta+\nu+2}. \tag{10}$$

Furthermore, we always apply the initial values

$$S(t) = 0.99, \quad I(t) = 0.01, \quad R(t) = 0 \qquad \text{for} \quad t \leq 0,$$

Table 2 Test cases in numerical simulation

Cases	Minimum τ_{min}	Maximum τ_{max}	Expected value μ
Case (i)	10	30	$16.\overline{6}$
Case (ii)	50	90	$63.\overline{3}$
Case (iii)	150	250	$183.\overline{3}$

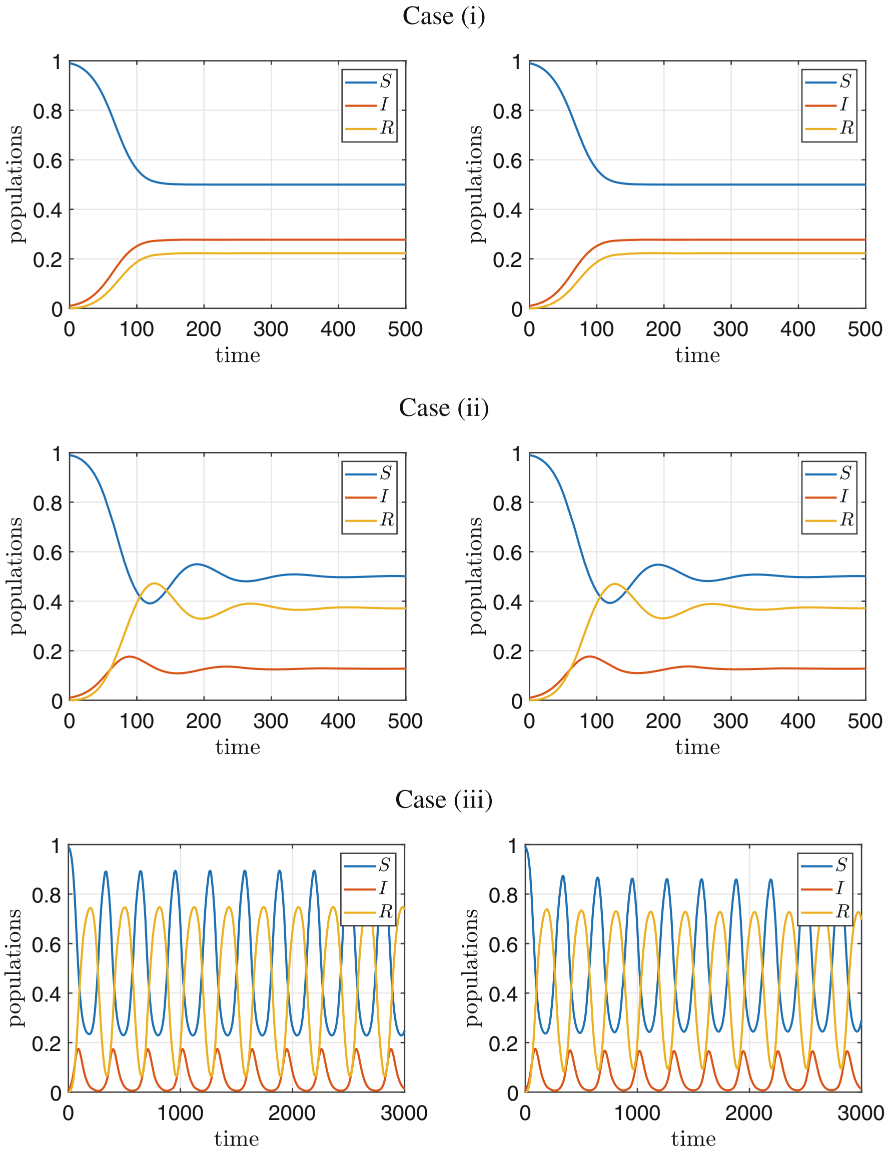
which are located close to a disease-free steady state of the form (3). The distributed DDEs (5) are solved by the numerical method in Sect. 4, where the Gauss-Jacobi quadrature is applied with $k = 8$ nodes. For comparison, we solve the DDEs (1) with the single delay $\tau = \mu$ from (10) using the above initial values.

The numerical computations were performed in the software package MATLAB, see [3]. We used the routine `dde23` to solve the DDEs with finite numbers of delays. Three test cases were examined, which are determined by the choice of $\tau_{\text{min}}, \tau_{\text{max}}$ in (9). Table 2 shows these cases.

Figure 1 illustrates the numerical solutions of the DDEs. In all three cases, the qualitative behaviour of the solutions coincides for the models (1) and (5). In case (i) and case (ii), the transient solutions converge to an endemic steady state of the form (4). This convergence is monotone in case (i), whereas damped oscillations occur in case (ii). Furthermore, the models (1) and (5) yield nearly the same solutions quantitatively. In case (iii), a Hopf bifurcation takes place and thus the solutions converge to a periodic steady state. Now the solutions of the models (1) and (5) exhibit small quantitative differences. We remark that quantitative differences increase if the delay $\tau = \mu$ in (1) is replaced by the centre $\tau = \frac{1}{2}(\tau_{\text{min}} + \tau_{\text{max}})$, for example.

6 Conclusions

We examined two SIR models including temporary immunity: a system of DDEs with a single delay and a system of distributed DDEs. Stationary solutions of the two models coincide. Numerical simulations show that the solutions of initial value problems exhibit the same qualitative properties in the models. Moreover, the solutions are often nearly identical provided that the single delay is chosen as the expected value of the probability distribution for the random delay.


Case (i)

Case (ii)

Case (iii)

Fig. 1 Population densities for the three test cases: model with single delay (left column) and model with distributed delay (right column)

References

1. F. Brauer and C. Castillo-Chavez (2001). *Mathematical models in population biology and epidemiology.* Springer, New York.
2. N.F. Britton (2003). *Essential Mathematical Biology.* Springer, London.
3. D.J. Higham and N.J. Higham (2017). *MATLAB Guide.* 3rd edn., SIAM, Philadelphia.
4. L.F. Shampine and S. Thompson (2009). Numerical solution of delay differential equations. in: *Delay Differential Equations: Recent Advances and New Directions.* Springer, Boston.
5. H. Smith (2011). *An Introduction to Delay Differential Equations with Applications to the Life Sciences.* Springer, New York.
6. J. Stoer and R. Bulirsch (2002). *Introduction to Numerical Analysis.* 3rd edn., Springer, New York.
7. T.J. Sullivan (2015). *Introduction to Uncertainty Quantification.* Springer, Switzerland.
8. M.L. Taylor and T.W. Carr (2009). An SIR epidemic model with partial temporary immunity modeled with delay. *J. Math. Biol.* 59:841–880.
9. D. Xiu (2010). *Numerical methods for stochastic computations: a spectral method approach.* Princeton University Press, New Jersey.

Next-Gen Gas Network Simulation

Christian Himpe, Sara Grundel, and Peter Benner

Abstract To overcome many-query optimization, control, or uncertainty quantification work loads in reliable gas and energy network operations, model order reduction is the mathematical technology of choice. To this end, we enhance the model, solver and reductor components of the `morgen` platform, introduced in Himpe et al. [J. Math. Ind. 11:13, 2021], and conclude with a mathematically, numerically and computationally favorable model-solver-reductor ensemble.

1 Model Order Reduction for Gas and Energy Networks

Computer-based simulation of gas transport in pipeline networks has been an industrial as well as academic field of interest since the earliest scientific computing systems [5]. Especially, the transient simulation of gas flow and the dynamic gas network behavior are the pinnacle discipline in this regard. The MATLAB-based `morgen`—Model Order Reduction for Gas and Energy Networks—platform[1] continues this research by providing a modular open-source software simulation stack for the comparison and benchmarking of models (discretizations), solvers (time steppers), and reductors (model reduction algorithms) [3]. Beyond selecting apposite simulator components or ranking model reduction methods, an overall goal is the acceleration of forward simulations, so that many-query tasks relying thereon, such as optimization, control or uncertainty quantification, benefit in terms

[1] See: https://git.io/morgen

C. Himpe (✉) · S. Grundel · P. Benner
Computational Methods in Systems and Control Theory Group, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
e-mail: himpe@mpi-magdeburg.mpg.de; grundel@mpi-magdeburg.mpg.de; benner@mpi-magdeburg.mpg.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_15

Table 1 Available models in morgen in version 1.1

Name	Identifier	port-Hamiltonian?	Reference
Midpoint discretization	`ode_mid`	No	[3, Sec. 2.4.1]
Endpoint discretization	`ode_end`	Yes	[3, Sec. 2.4.2]

Table 2 Available solvers in morgen in version 1.1

Name	Identifier	Comment	Reference
Adaptive second order Rosenbrock	`generic`	Uses `ode23s`	[3, Sec. 5.3.1]
First order implicit-explicit	`imex1`	Non-Runge-Kutta	[3, Sec. 5.3.3]
Second order implicit-explicit	`imex2`	Runge-Kutta	[3, Sec. 5.3.4]
Explicit fourth order Runge-Kutta	`rk4`		[3, Sec. 5.3.2]
Explicit second order Runge-Kutta	`rk2hyp`	Increased stability	[9]
Explicit fourth order Runge-Kutta	`rh4hyp`	Increased stability	[6]

of performance. In this work, we summarize and enhance the foundational work of [3] with additional details, and accompany version 1.1 of morgen.

1.1 Modules Overview

The morgen platform is organized into modules: *models*, *solvers*, *reductors*, *networks* and *tests*. The *networks* module holds topology and scenario data, and the *tests* module defines the simulation and model reduction experiments, thus, we summarize the currently available core modules: *models*, *solvers*, and *reductors*. The *models* module assembles a semi-discrete input-output system from a network topology. Currently, two spatially discrete models are included (Table 1). The *solvers* module computes a time-discrete output trajectory from a model and a scenario. Six solvers are provided in the current version (Table 2). The *reductors* module compresses a model given a solver and (generic training) scenario. All in all, 23 reductors organized in four classes are available (Table 3).

2 Enhanced Functionality

In this section, we discuss some of the new properties of the morgen 1.1 platform. Specifically, one aspect of each core module (*model*, *solver*, *reductor*) is addressed.

Table 3 Available reductors in morgen in version 1.1

Name	Identifier	Linear variant	Reference
Structured proper orthogonal decomposition	pod_r	–	[3, Sec. 4.2]
Structured empirical dominant subspaces	eds_ro	eds_ro_l	[3, Sec. 4.3]
Structured empirical dominant subspaces	eds_wx	eds_wx_l	[3, Sec. 4.3]
Structured empirical dominant subspaces	eds_wz	eds_wz_l	[3, Sec. 4.3]
Structured balanced POD	bpod_ro	bpod_ro_l	[3, Sec. 4.4.3]
Structured balanced truncation	ebt_ro	ebt_ro_l	[3, Sec. 4.4]
Structured balanced truncation	ebt_wx	ebt_wx_l	[3, Sec. 4.4]
Structured balanced truncation	ebt_wz	ebt_wz_l	[3, Sec. 4.4]
Structured goal-oriented POD	gopod_r	–	[3, Sec. 4.5.1]
Structured balanced gains	ebg_ro	ebg_ro_l	[3, Sec. 4.5]
Structured balanced gains	ebg_wx	ebg_wx_l	[3, Sec. 4.5]
Structured balanced gains	ebg_wz	ebg_wz_l	[3, Sec. 4.5]
Structured DMD Galerkin	dmd_r	–	[3, Sec. 4.6]

2.1 Gravity Term

One component of the gas pipeline model, particularly of the retarding forces in the mass-flux equation, is the gravity term, which accounts for increase or decrease in momentum due to an incline in a pipeline section. In [2], this gravity term is modeled in great detail, as it does not only consider a height difference between the pipe's end points, as morgen does, but also the height profile for the full run of the pipe (see [2, Fig. 11]). Both approaches are justified, depending on the aimed accuracy of the model, as discussed in [1]. Such pipeline height profiles can be included into morgen by supplying a pipe as sequence of virtual pipes, each connecting two subsequent local height extrema. In morgen 1.1, the gravity term is configurable so it is computable based on the dynamic pressure, static pressure or not at all, whereas the static gravity term, based on the steady-state was newly added.

2.2 Explicit Solvers

In [3], the classic explicit 4th order Runge-Kutta method rk4 was tested, as it was employed in earlier works. Yet we found it to be *not* suitable for gas network simulations. In [4], an explicit Runge-Kutta method from [9, Sec. 4] was suggested for this application. The Butcher tableau for this explicit 5-stage, 2nd order low-storage scheme with increased stability, is given by:

$$\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{4} & \frac{1}{4} & & & & \\
\frac{1}{6} & 0 & \frac{1}{6} & & & \\
\frac{3}{8} & 0 & 0 & \frac{3}{8} & & \\
\frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & \\
\hline
 & 0 & 0 & 0 & 0 & 1
\end{array}$$

This additional solver `rk2hyp`, as well as a 4th order Runge-Kutta method with increased *hyperbolic stability limit* from [6, Sec. 4.1] (`rk4hyp`), were added to `morgen` 1.1 and tested against various test problems. Both increased-stability solvers allow larger time steps then `rk4`, specifically in conjunction with the `ode_end` model, but compared to the implicit-explicit solvers `imex1` and `imex2`, they are still not fully competitive. However, these explicit methods could be interesting for new implicit-explicit or predictor-corrector methods.

2.3 Gain Matching

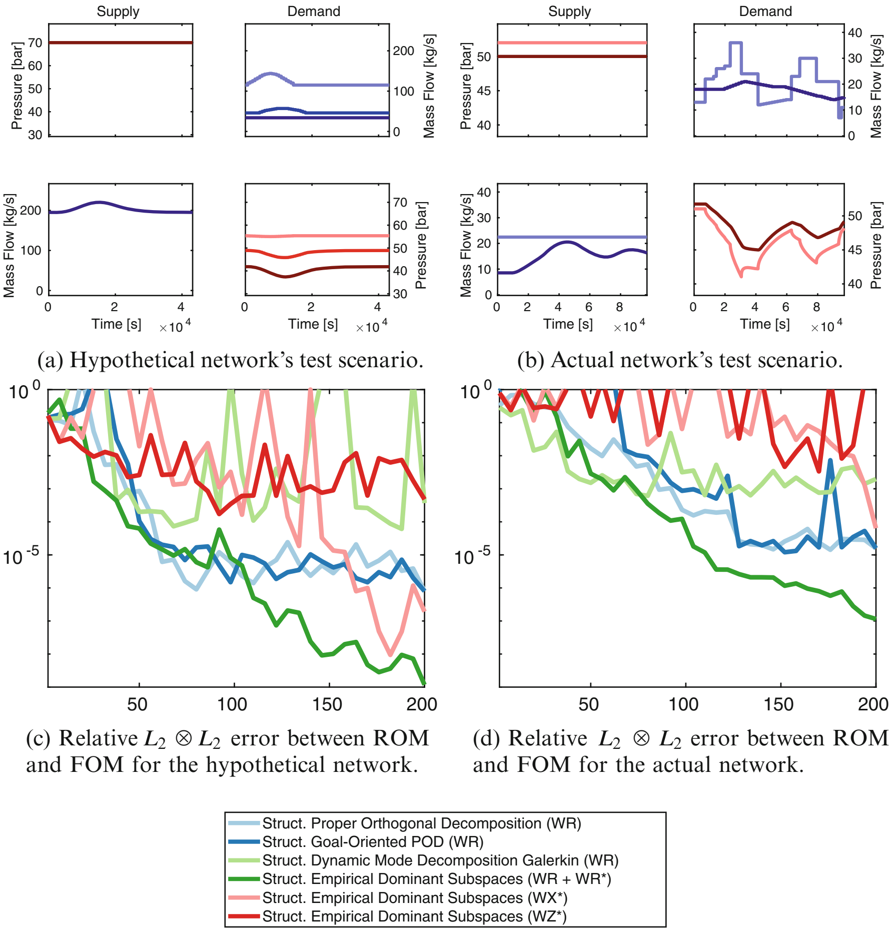
An important quality for certain applications of model reduction, such as electrical circuits, is the preservation of the steady-state gain (also known as DC gain), which is the output for zero frequency input. First, we clarify that we are not discussing the actual steady-state gain of the reduced order model, due to the centering around the steady-state and hence, the steady-state gain match [3, Sec. 3]. Yet, there can still be an output error for a constant input on top of the steady-state input, which is relevant due to the assumed low-frequency boundary values. Since there is an interpretation of gas networks as circuits [8], we consider this reduced model property, which induces two questions: How to compute the steady-state gain, and how to correct a gain mismatch? The former is answered by [10], stating that for a linear port-Hamiltonian model, with components as in [3, Sec. 2.9], the gain S is computable by:

$$S = CQ^{-1}B,$$

with input matrix B, output matrix C, and energy storage matrix Q. Since the models are nonlinear and do not have to be port-Hamiltonian, but comprise the same model components, the above formula can still be applied albeit yielding only an approximation. The per-port gain mismatch D_* is then computed by the difference of full and reduced-order model (reduced-order quantities are denoted by $\cdot_r$) gain:

$$D_* := (CQ^{-1}B) - (C_r Q_r^{-1} B_r),$$

which can then be used to correct the reduced-order model gain by adding D_* as a feedthrough matrix to the output function, as described in the gain matching procedure in [7]. We included this approximate gain matching test to `morgen` 1.1.



(a) Hypothetical network's test scenario.

(b) Actual network's test scenario.

(c) Relative $L_2 \otimes L_2$ error between ROM and FOM for the hypothetical network.

(d) Relative $L_2 \otimes L_2$ error between ROM and FOM for the actual network.

(e) Common legend for the model reduction error plots.

Reductor	MORScore	**Avg. Gain Error**
`pod_r`	0.27	$6 \cdot 10^{-6}$
`gopod_r`	0.26	$6 \cdot 10^{-6}$
`dmd_r`	0.18	$8 \cdot 10^{-6}$
`eds_ro_l`	0.30	$8 \cdot 10^{-6}$
`eds_wx_l`	0.18	$8 \cdot 10^{-6}$
`eds_wz_l`	0.15	$8 \cdot 10^{-6}$

Reductor	MORScore	**Avg. Gain Error**
`pod_r`	0.19	$2 \cdot 10^{-5}$
`gopod_r`	0.15	$1 \cdot 10^{-5}$
`dmd_r`	0.15	$2 \cdot 10^{-5}$
`eds_ro_l`	0.24	$2 \cdot 10^{-5}$
`eds_wx_l`	0.04	$2 \cdot 10^{-5}$
`eds_wz_l`	0.03	$2 \cdot 10^{-5}$

(f) MORScore $\mu(200, \epsilon_{\text{mach}(16)}) \in (0, 1]$ in the $L_2 \otimes L_2$ error norm (higher means more accurate ROM), and mean steady-state gain error for the hypothetical network.

(g) MORScore $\mu(200, \epsilon_{\text{mach}(16)}) \in (0, 1]$ in the $L_2 \otimes L_2$ error norm (higher means more accurate ROM), and mean steady-state gain error for the actual network.

Fig. 1 Visualization of the test scenario, model reduction errors between FOM (full-order model) and ROM (reduced-order model), MORScore, and gain errors of the tested ROMs for the hypothetical network [5, Part 2] (left side) and actual network [5, Part 3] (right side). Computed with MATLAB 2021a. See [3, Sec. 6] for a description of the plot presentation

The gain correction was tested with all reductors (Table 3). For all reductors, the correction was about the level of 10^{-5}, see Tables f and g in Fig. 1, except for the `bpod_ro` method, for which the gain correction fully deteriorates the reduced model. Thus, the improvement of reduced-order models is small at best. This is not unexpected, considering the gas network model is hyperbolic: A single pipeline, or more generally an input-output system based on a first-order hyperbolic partial differential equation, has the transport property which expresses as a delay in observable outputs of controllable inputs. Hence, an immediate effect of inputs to outputs (circumventing the system dynamics), i.e. by a feedthrough term, is typically not needed.

3 Numerical Experiments

We extend the numerical experiments in [3], by reimplementing the results from [5], specifically, we test the hypothetical network [5, Part 2], and the actual network [5, Part 3], which are both tree networks, on their associated scenarios.

Six structured empirical-Gramian-based Galerkin reductors are tested on the port-Hamiltonian endpoint model and the first order implicit-explicit solver. The results are presented in Fig. 1. In line with other experiments, the `eds_ro_l` reductor yields the most accurate results.

4 Next-Gen Gas Network Simulation

For the newly tested features we conclude that currently, explicit solvers do not seem a viable option to simulate gas networks, while gain matching offers only minor accuracy improvements; yet, the new static gravity term is more robust with respect to model reduction and is henceforth the default setting in `morgen`.

Overall, based on the comparisons in [3] and this work's numerical results, we currently recommend a port-Hamiltonian model, an implicit-explicit solver, and a Galerkin reductor. Thus, the endpoint discretization, first order IMEX time stepper, and the structured empirical dominant subspaces reductor make a promising model-solver-reductor ensemble for the next generation of transient gas network simulators. Future extensions of the `morgen` platform will refine this recommendation.

References

1. G. Bachman and M. Goodreau. Less is more accuracy versus precision in modeling. In PSIG Annual Meeting 2000, pp. PSIG–0009, 2000.

2. M. Behbahani-Nejad, A. Bermúdez, and M. Shabani. Finite element solution of a new formulation for gas flow in a pipe with source terms. J. Natural Gas Sci. Engrg. 61:237–250, 2019.
3. C. Himpe, S. Grundel, and P. Benner. Model order reduction for gas and energy networks. J. Math. Ind., 11:13, 2021.
4. A. Lewandowski. New numerical methods for transient modeling of gas pipeline networks. In PSIG Annual Meeting, pp. PSIG–9510, 1995.
5. L.A. Lotito and P.W. Halbert. Computer simulation of gas flow dynamics. Pipeline Engineer, 39:31–33, 29–31, 45–47, 1967.
6. J.L. Mead and R.A. Renaut. Optimal Runge-Kutta methods for first order pseudospectral operators. J. Comput. Phys. 152(1):404–419, 1999.
7. R. Samar, I. Postlewaite, and D.W. Gu. Model reduction with balanced realizations. Int. J. Control, 62(1):33–64, 1995.
8. W.Q. Tao and H.C. Ti. Transient analysis of gas pipeline network. Chem. Engrg. J., 69(1):47–52, 1998.
9. P.J. van der Houwen. Explicit Runge-Kutta formulas with increased stability boundaries. Numer. Math. 20:149–164, 1972.
10. A. van der Schaft. Interconnections of input-output Hamiltonian systems with dissipation. In Proceedings of the 55th IEEE Conference on Decision and Control, pp. 4886–4691, 2016.

Parameter Estimation via Adjoint Functions in Epidemiological Reaction-Diffusion Models

Peter Heidrich and Thomas Götz

Abstract The current pandemic situation due to COVID-19 demonstrates the need for epidemiologic models to represent infection events as accurately as possible. An important factor is the mobility of the affected individuals which can be investigated with discrete or continuous spatial models. In this contribution, parameter estimation via adjoint functions is presented to fit a reaction-diffusion PDE system with epidemiological SIS model to data sets. For this purpose static and dynamic optimization methods are used to solve an L^2-norm based least squares problem. An artificial data set is generated to test the accuracy of the procedure. Subsequently, the PDE system is adapted to this data set using methods of optimal control theory. Unknown parameters like diffusivity and transmission rate can be determined. The noise in the data set is also taken into account by fitting the initial conditions. The results show that the method is well suited for this purpose and should be further used with real data sets.

1 Introduction

The example of the current COVID-19 pandemic clearly shows the significant influence of mobility on the spread of a disease. Mathematical-epidemiological models can address this using various techniques. The movement of people between separate patches such as airports, islands, cities etc. can be represented using Lagrangian movement for short-term stays or Eulerian movement for long-term

P. Heidrich (✉)
Mathematical Institute, University of Koblenz-Landau, Campus Koblenz, Koblenz, Germany

Magister Laukhard IGS Herrstein-Rhaunen, Herrstein, Germany
e-mail: heidrich@uni-koblenz.de

T. Götz
Mathematical Institute, University of Koblenz-Landau, Campus Koblenz, Koblenz, Germany
e-mail: goetz@uni-koblenz.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_16

migrations [5]. The modelling here is done via ordinary differential equations (ODE). However, since this point-by-point distribution of pathogens does not reflect reality on its own, spatial spread need to be taken into account as well. This can be achieved with reaction-diffusion systems which contain partial differential equations (PDE) [2, 10]. Consequently, we consider a system of the form

$$\begin{aligned} \partial_t u &= \kappa \Delta u + f(u)\,, \\ u &= u_0, \quad t = 0\,, \\ \partial_\nu u &= 0, \quad x \in \partial\Omega\,. \end{aligned}$$

The goal is to fit this model to data sets. Unfortunately, several parameters are unknown in the epidemiological context, such as the transmission rate or even the parameters describing mobility. In addition, noisy data may be expected, for this the initial value condition shall be adjusted.

In this contribution, a parameter estimation via adjoint functions is tested. This corresponds to techniques from static and dynamic optimization. To investigate the accuracy of the method, we consider an artificially generated data set. Numerical simulations are performed to fit the model to this data set.

2 Model

In the following we consider the set $\Omega = (0, a) \times (0, b)$ as spatial coordinates and a time axis $(0, T)$ with resulting domain $V = \Omega \times (0, T)$. To model a spatial spread of an infectious disease, we use an epidemiological SIS model. The resulting reaction-diffusion system reads as

$$\partial_t S = \kappa_S \Delta S - \frac{\beta}{N} S I + \gamma I\,, \tag{1a}$$

$$\partial_t I = \kappa_I \Delta I + \frac{\beta}{N} S I - \gamma I\,, \tag{1b}$$

$$S = S_0,\; I = I_0, \quad t = 0\,, \tag{1c}$$

$$\partial_\nu S = \partial_\nu I = 0, \quad x \in \partial\Omega\,. \tag{1d}$$

The functions $S, I, N \in C^{2,1}(V)$ represent the densities of the compartments of susceptible (S) and infected (I) individuals and the total population density $N = S + I$ in coordinate x at time t.

Here, e.g. $\partial_t S = \frac{\partial S}{\partial t}$ stands for the time derivative of S and $\Delta S = \operatorname{div}(\operatorname{grad} S) = \frac{\partial^2 S}{\partial x_1^2} + \frac{\partial^2 S}{\partial x_2^2}$ stands for the Laplace operator. For the two compartments initial value conditions are given by $S_0, I_0 \in C^2(\Omega)$. At the boundary $\partial\Omega$ Neumann boundary conditions are implied, whereby $\partial_\nu S$ denotes the derivative of S to the direction of the unit outward normal ν. In context, this means that none of the individuals leaves the region Ω. We also assume that $\int_\Omega I(x,0)\,dx > 0$ holds with $S_0, I_0 \geq 0$. We define

$$\overline{N} := \int_\Omega N(x,0)\,dx\,,$$

which stands for the total number of individuals at time $t = 0$. Due to the Neumann boundary conditions the Gauss's theorem delivers

$$\frac{\partial}{\partial t}\int_\Omega N(x,t)\,dx = \int_\Omega \kappa_S \Delta S + \kappa_I \Delta I\; dx = \int_{\partial\Omega} \kappa_S \partial_\nu S + \kappa_I \partial_\nu I\; d\omega = 0\,.$$

Thus, the total population is constant with value $\overline{N}$.

The parameters $\beta, \gamma > 0$ represent the transmission and recovery rates of the corresponding disease and $\kappa_S, \kappa_I > 0$ the diffusivity of the corresponding compartments. For simplicity, we assume that $\kappa_s = \kappa_I$ holds and β, γ are constants independent of x. For the derivation of such a model in one dimensional case and the operation of epidemiological models, we refer to [5].

SIS-based reaction-diffusion systems as in (1) have already been studied in [1, 3, 6–9]. The existence of a global and unique solution is shown, also for cases in which $\kappa_S \neq \kappa_I$ holds and β, γ are Hölder continuous functions over Ω. In [1] a Basic Reproduction Number is established on Sobolev space $H^1(\Omega)$ by

$$\mathcal{R}_0 = \sup_{\substack{\varphi \in H^1(\Omega)\\ \varphi \neq 0}} \left(\frac{\int_\Omega \beta\varphi^2}{\int_\Omega \kappa_I |\nabla\varphi|^2 + \gamma\varphi^2} \right). \tag{2}$$

There is shown, that if $\mathcal{R}_0 < 1$ holds, the unique disease-free equilibrium DFE $= \left(\frac{\overline{N}}{|\Omega|}, 0\right)$ is globally asymptotically stable and unstable for $\mathcal{R}_0 > 1$. The expression $|\Omega|$ here stands for the corresponding measure. On the other hand, for $\mathcal{R}_0 > 1$ the existence of a unique endemic equilibrium EE is shown.

Furthermore, we set $\kappa := \kappa_S = \kappa_I$ and substitute $S = N - I$. If we additionally define $u := \frac{I}{N}$, we receive a reduced system with $f(u) := \beta(1-u)u - \gamma u$

$$\partial_t u = \kappa \Delta u + f(u)\,, \tag{3a}$$

$$u = u_0, \quad t = 0\,, \tag{3b}$$

$$\partial_\nu u = 0, \quad x \in \partial\Omega\,. \tag{3c}$$

The simplifying assumptions and the normalization are used to test the presented parameter fitting via adjoint functions. It is clear that in realistic situations much more complex models should be used.

3 Adjoint System

We now want to fit model (3) to data sets using adjoint functions known from optimal control theory. In the epidemiological context, this means parameter estimation of the transmission rate $\beta > 0$ and diffusivity $\kappa > 0$. The recovery rate $\gamma > 0$ can be assumed to be the reciprocal of the average infection duration and thus does not need to be fitted. Furthermore, we assume that the data is noisy and therefore the initial condition $u_0 \in C^2(\Omega)$ has to be adjusted. In the following, the function u^{DATA} contains the available data points and u_0^{DATA} the supposedly noisy initial value of the data set at $t = 0$.

We introduce an objective function $J : \mathbb{R}^2 \times C^2(\Omega) \to \mathbb{R}$

$$J(\beta, \kappa, u_0) := w_0 \|u - u^{DATA}\|^2_{L^2_V} + w_1(\beta^2 + \kappa^2) + w_2 \|u_0 - u_0^{DATA}\|^2_{L^2_\Omega} \,. \tag{4}$$

The function u stands for the solution of the reaction-diffusion PDE system (3). The objective function includes the L^2-norm $\|g\|_{L^2_Y} := \left(\int_Y g(y)^2\, dy\right)^{1/2}$ and corresponding normalizing weights $w_0 := 1/\|u^{DATA}\|^2_{L^2_V}$ respectively $w_2 := 1/\|u_0^{DATA}\|^2_{L^2_\Omega}$. The convex and radially unbounded regularization term $w_1(\beta^2 + \kappa^2)$ depends on a very small choosen weight w_1 whose influence is investigated in the subsequent simulations. Assuming one already has initial guess $\hat{\beta}, \hat{\kappa}$ for the parameters, a term of the form $w_1 \left((\beta - \hat{\beta})^2 + (\kappa - \hat{\kappa})^2\right)$ can be used alternatively.

This leads to a minimization problem with dynamic constraints

$$\min_{\beta,\kappa,u_0} J(\beta, \kappa, u_0) \qquad \text{subject to PDE system (3)} \,. \tag{5}$$

A Lagrange function is introduced containing adjoint functions $z \in C^{2,1}(V)$

$$L(\beta, \kappa, u_0, u, z) := \int_V g\, dxdt + \psi + \int_V z\, (f(u) + \kappa \Delta u - \partial_t u)\, dxdt \,, \tag{6}$$

whereby $g := w_0 \left(u - u^{DATA}\right)^2$ and $\psi := w_1(\beta^2 + \kappa^2) + w_2 \int_\Omega \left(u_0 - u_0^{DATA}\right)^2 dx$.

The necessary condition for a minimum $(\beta^*, \kappa^*, u_0^*, u^*, z^*)$ is fulfilled, if

$$0 = \nabla L := \left(\partial_\beta L, \partial_\kappa L, \partial_{u_0} L, \partial_u L, \partial_z L\right)$$

holds true. It should be noted that Gâteaux derivatives are needed for the derivatives of L to the directions u_0, u and z. This leads to the following system in $(\beta^*, \kappa^*, u_0^*, u^*, z^*)$:

(i) $0 = \partial_\beta \psi + \int_V z \partial_\beta f \, dxdt,$ (Optimality Condition)
$0 = \partial_\kappa \psi + \int_V z \Delta u \, dxdt,$
(ii) $u_0 = u_0^{DATA} - \frac{z(x,0)}{2w_2},$ (Optimal Initial Condition)
(iii) $\partial_t z = -\partial_u g - z \partial_u f - \kappa \Delta z,$ (Adjoint Equation)
$z = 0, \quad t = T,$ (Transversality Condition)
$\partial_\nu z = 0, \quad x \in \partial\Omega,$ (Adjoint Neumann Boundary Condition).

When L is derived in the z direction, the original PDE system (3) is recovered.

4 Numerical Simulations

From the analysis in Sect. 3, the gradient of L with respect to β and κ reads

$$\partial_\beta L = 2w_1 \beta + \int_V z(1-u)u \, dxdt \tag{7a}$$

$$\partial_\kappa L = 2w_1 \kappa + \int_V z \Delta u \, dxdt \tag{7b}$$

and we obtain the adjoint equation

$$\partial_t z = -2w_0 \left(u - u^{DATA} \right) - z(\beta(1-2u) - \gamma) - \kappa \Delta z \,. \tag{8}$$

The latter must be solved backward in time t due to the transversality condition. This is done using the forward-backward sweep method, see [4]. The performed algorithm can be found in Appendix 1. Solving the PDEs is done using finite differences

$$\Delta u_{i,j}^n \approx \frac{1}{h^2} \left(u_{i-1,j}^n + u_{i,j-1}^n - 4u_{i,j}^n + u_{i+1,j}^n + u_{i,j+1}^n \right) \tag{9}$$

and an explicit Euler-scheme

$$u_{i,j}^{n+1} = u_{i,j}^n + \tau(\kappa \Delta u_{i,j}^n + f(u_{i,j}^n)) \tag{10}$$

on the domain $V = \Omega \times (0, T)$ with $\Omega = (0, a) \times (0, b)$. The Neumann boundary conditions are implemented by $u_{k+1,j}^n = u_{k,j}^n$ etc., if index (k, j) stands for a point at the rectangular boundary $\partial\Omega$. In the following simulations we use the setting

- $h := 0.1, \tau := 0.001, a := 3, b := 2, T := 1$
- $x_1^i = ih$: $i = 0, \ldots, 30$ $x_2^j = jh$: $j = 0, \ldots, 20$ $t^n = n\tau$: $n = 0, \ldots, 1000$.

To test the procedure an artificial data set is generated with initial condition

$$u_0^{DATA}(x_1^i, x_2^j) := 0.02\delta_{(0.4,0.6)}(x_1^i, x_2^j) + 0.1\delta_{(2,1)}(x_1^i, x_2^j) \tag{11}$$

whereby $\delta_{(\tilde{x}_1,\tilde{x}_2)}(x_1^i, x_2^j) = 1$, if $(x_1^i, x_2^j) = (\tilde{x}_1, \tilde{x}_2)$ and else $\delta_{(\tilde{x}_1,\tilde{x}_2)} = 0$. Subsequently, the state variable PDE (3) is solved with $\beta := 0.3$, $\kappa := 0.2$ and $\gamma := 0.1$. The received solution is called $\overline{u}$ in the following. To simulate noisy data, a normally distributed $q_{i,j}^n \sim \mathcal{N}(0, \sigma^2)$ is generated, so that the desired data set is calculated by

$$u^{DATA}(x_1^i, x_2^j, t^n) := \max\Big(0, (1 + q_{i,j}^n) \cdot \overline{u}(x_1^i, x_2^j, t^n)\Big)\,. \tag{12}$$

5 Results and Conclusions

The application of the presented method is tested in three simulations with different initial values β_0, κ_0. The initial value for the initial condition u_0 is taken from the desired data set u^{DATA}. The resulting Table 1 and Fig. 1 in Appendix 2 show adequate parameter estimates. A test run without artificial noise on the data set resulted in the original values $\beta = 0.3$ and $\kappa = 0.2$. The simulations also show the effect of the weight w_1 of the regularization term on the minimization of the objective function J. Despite this disturbance, better results are obtained than without it. The prerequisite for this is a correspondingly small choice for w_1 which influences the convexity of the objective function in the respective parameters.

The present simulations show that the applied method works very well in this toy problem with self-generated data set. In principle, the procedure is suitable to perform such parameter estimations. In the next step, the method should be tested with real data sets. Depending on the disease, much more sophisticated epidemiological models may also be required. Mobility movements between patches, such as daily commuting or travelling, should also be added to the model. With respect to the PDE solution, other solution methods should also be tested, since the simple Euler method may be numerically unstable. In addition, a simple rectangular area was assumed in our example. In real cases, appropriate adjustments are necessary here.

Appendix 1

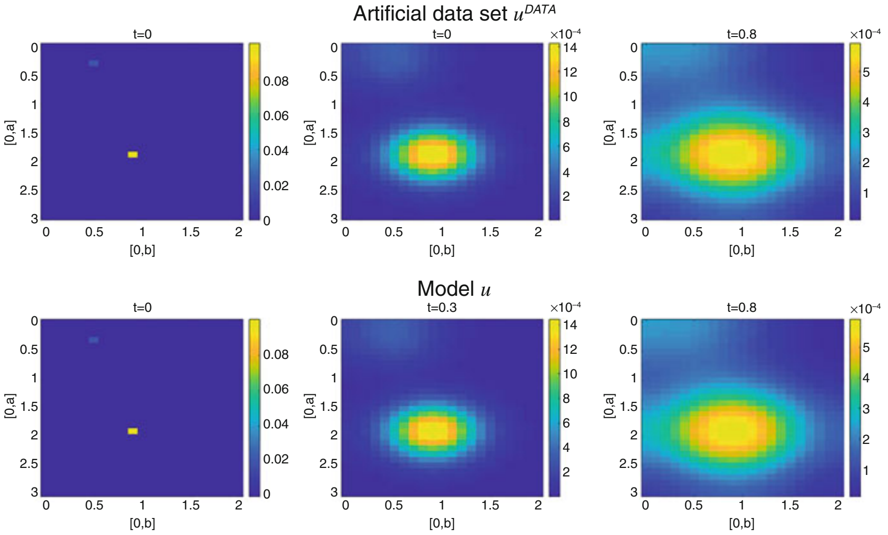
Algorithm 1 Pseudocode for the parameter estimation via adjoint functions

```
$\beta, \kappa, u^{DATA}, u_0^{DATA}$ ← load initial values and data
$u, z$ ← solve PDE for state variable and adjoint function
$J,\ \nabla J$ ← compute objective function and gradient regarding$\beta$ and $\kappa$
$s_1$ ← compute search direction for$\beta$ and $\kappa$(Quasi-Newton (BFGS))
$s_2 \leftarrow (\tilde{u}_0 - u_0)$ compute search direction for $u_0$ with $\tilde{u}_0 = u_0^{DATA} - \frac{z(x,0)}{2w_2}$
repeat
    $J_{old} \leftarrow J$
    $\theta \leftarrow 1$
    $(\beta, \kappa) \leftarrow (\beta, \kappa) + \theta s_1$
    $u_0 \leftarrow u_0 + \theta s_2$
    $u, J$ ← update
    repeat
        $\theta \leftarrow 0.5\theta$
        $(\beta, \kappa) \leftarrow (\beta, \kappa) + \theta s_1$
        $u_0 \leftarrow u_0 + \theta s_2$
        $u, J$ ← update
    until $J \leq J_{old} + 0.001\theta s^T \nabla J_{old}$   (Armijo Rule)
    $z, \nabla J, s_1, s_2$ ← update
until $\frac{\|J - J_{old}\|_2}{\|J_{old}\|_2} < \text{TOL}$
```

Appendix 2

Table 1 The recovery rate is fixed with $\gamma := 0.1$. The algorithm stops with tolerance TOL $:= 10^{-6}$. The original parameters of the artificial data set are $\beta := 0.3$ and $\kappa := 0.2$. The artificial noise is generated with standard deviation $\sigma := 0.1$

Simulation	β	κ	J	w_1	Iterations
$\beta_0 := 0.5, \kappa_0 := 0.4$					
Best fit	0.2796	0.1994	$1.16 \cdot 10^{-4}$	10^{-07}	242
	0.2787	0.1994	$1.17 \cdot 10^{-4}$	10^{-08}	364
	0.2496	0.1988	$1.65 \cdot 10^{-4}$	10^{-09}	197
	0.2609	0.1990	$1.42 \cdot 10^{-4}$	0	240
$\beta_0 := 0.1, \kappa_0 := 0.5$					
Best fit	0.2659	0.1991	$1.33 \cdot 10^{-4}$	10^{-08}	181
	0.2251	0.1983	$2.36 \cdot 10^{-4}$	10^{-09}	364
	0.2857	0.1995	$1.11 \cdot 10^{-4}$	10^{-10}	345
	0.2753	0.1993	$1.20 \cdot 10^{-4}$	0	388
$\beta_0 := 1.0, \kappa_0 := 1.0$					
Best fit	0.2499	0.1988	$1.64 \cdot 10^{-4}$	10^{-10}	314
	0.2824	0.1994	$1.14 \cdot 10^{-4}$	10^{-11}	655
	0.2836	0.1994	$1.11 \cdot 10^{-4}$	10^{-12}	555
	0.2540	0.1988	$1.55 \cdot 10^{-4}$	0	304


Artificial data set u^{DATA}

Model u

Fig. 1 Graphical Results for the Simulation with $\beta_0 := 1$, $\kappa_0 := 1$, $w_1 := 10^{-12}$

References

1. Allen, L.J.S., Bolker, B.M., Lou, Y., Nevai, A.L.: Asymptotic profiles of the steady states for an SIS epidemic reaction–diffusion model. Discrete Contin. Dynam. Syst., **21**, 1–20 (2008)
2. Britton, N.F.: Reaction-Diffusion Equations and Their Applications to Biology. Academic Press, London (1986)
3. Deng, K., Wu, Yixiang: Dynamics of a susceptible–infected–susceptible epidemic reaction–diffusion model. Proceedings of the Royal Society of Edinburgh: Section A Mathematics, **146(5)**, 929–946 (2016)
4. Lenhart, S., Workman, J.T.: Optimal control applied to biological models. CRC Press, 2007.
5. Martcheva, M.: An Introduction to Mathematical Epidemiology. Springer, New York (2015)
6. Peng, R.: Asymptotic profiles of the positive steady state for an SIS epidemic reaction–diffusion model. J. Diff. Eqns, **247**, 1096–1119 (2009)
7. Peng, R., Liu, S.: Global stability of the steady states of an SIS epidemic reaction–diffusion model. Nonlin. Analysis, **71**, 239–247 (2009)
8. Peng, R., Yi, F.: Asymptotic profile of the positive steady state for an SIS epidemic reaction–diffusion model: effects of epidemic risk and population movement. Physica D, **259**, 8–25 (2013)
9. Peng, R., Zhao, X.: A reaction–diffusion SIS epidemic model in a time-periodic environment. Nonlinearity, **25**, 1451–1471 (2012)
10. Webb, G.F.: A Reaction-Diffusion Model for a Deterministic Diffusive Epidemic. J. Math. Anal. Appl., **84**, 150–161 (1986)

Graph-Based View of an Equilibrium Model for Nonwoven Tensile Strength Simulations

Marc Harmening, Nicole Marheineke, and Raimund Wegener

Abstract Focus of this work is the graph-based analytical treatment of the equilibrium model introduced in [4], which allows to determine the tensile behavior of nonwovens over the interaction of the individual fiber connections in the material. We use the representation of fiber structures as arbitrarily directed graphs to derive a compact nonlinear system of equations with characteristic divergence structure and to investigate its solvability and the uniqueness of solution. Further, we discuss the identification of subgraphs for which trivial solutions can be found.

1 Equilibrium Model

The microstructure of nonwovens consists of thousands of fibers bonded, for example, by thermal or chemical means. Their topology can be described by arbitrarily oriented graphs $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where the nodes $\mathcal{N}$ represent both adhesive joints and fiber ends, and the edges $\mathcal{E}$ represent the individual fiber connections between them (see Fig. 1). The spatial positions of the adhesive joints and fiber ends (nodes) are denoted by $\mathbf{x} \in \mathbb{R}^{3|\mathcal{N}|}$. To refer to the position of an individual node $\nu \in \mathcal{N}$ we write $\mathbf{x}_\nu \in \mathbb{R}^3$. Similarly, $\ell_\mu \in \mathbb{R}_+$ refers to the (positive) length of the fiber connection represented by edge $\mu \in \mathcal{E}$, yielding a global length vector $\ell \in \mathbb{R}_+^{|\mathcal{E}|}$.

To model the nonwoven tensile behavior, we consider the truss-based approach introduced in [4]. Thus, we distinguish further between boundary nodes $\mathcal{N}_B$ and interior nodes $\mathcal{N}_I$, such that $\mathcal{N} = \mathcal{N}_I \mathbin{\dot{\cup}} \mathcal{N}_B$. Thereby, the positions of the boundary nodes are fixed, while the positions of the remaining interior nodes are determined

M. Harmening (✉) · N. Marheineke
Trier University, Trier, Germany
e-mail: harmening@uni-trier.de

R. Wegener
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), Kaiserslautern, Germany

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_17

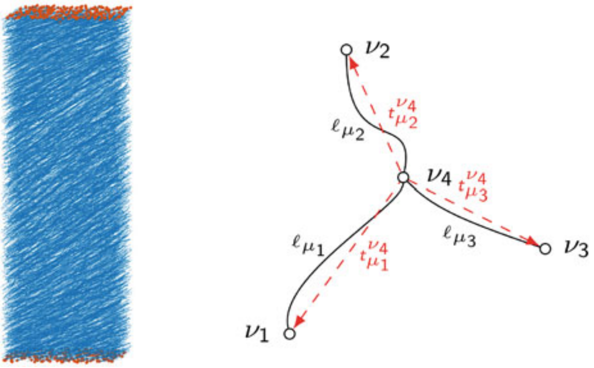


Fig. 1 Graph representation. Left: Topology of a virtually generated nonwoven material sample, cf. [4], with boundary nodes highlighted in red. Right: Simple fiber structure constellation, where the black lines mark fiber connections between the nodes and the dashed red lines indicate the corresponding edges representing the fiber connections

by a force equilibrium condition that accounts for the static material behavior. For the forces acting on the interior nodes, the model is restricted to the stresses caused by strain on incident fiber connections, which results in the following system:

$$\mathbf{x}_\nu = \mathbf{g}_\nu, \qquad \forall \nu \in \mathcal{N}_B, \tag{1}$$

$$\sum_{\mu \in \mathcal{E}(\nu)} \mathbf{f}^\nu_\mu(\mathbf{x}) = \mathbf{0}, \quad \mathbf{f}^\nu_\mu(\mathbf{x}) = \frac{\mathbf{t}^\nu_\mu(\mathbf{x})}{\|\mathbf{t}^\nu_\mu(\mathbf{x})\|_2} N(\epsilon(\|\mathbf{t}^\nu_\mu(\mathbf{x})\|_2, \ell_\mu)), \quad \forall \nu \in \mathcal{N}_I, \tag{2}$$

where $\mathbf{g}_\nu \in \mathbb{R}^3$ is the position specified for node $\nu \in \mathcal{N}_B$, the set $\mathcal{E}(\nu) \subset \mathcal{E}$ consists of all edges incident to node ν and $\mathbf{f}^\nu_\mu : \mathbb{R}^{3|\mathcal{N}|} \to \mathbb{R}^3$ expresses the force acting on node $\nu \in \mathcal{N}_I$ which is caused by stress on edge $\mu \in \mathcal{E}(\nu)$. According to (2), we have that $\mathbf{f}^\nu_\mu$ acts in the normalized direction $\mathbf{t}^\nu_\mu(\mathbf{x}) = \mathbf{x}_{\tilde{\nu}} - \mathbf{x}_\nu$ for $\mu = (\nu, \tilde{\nu})$, where the amplitude $N : [-1, \infty[\to \mathbb{R}_+$ depends on the relative strain of the fiber connection with respect to its length ℓ_μ, i.e., $\epsilon : \mathbb{R}^+ \times \mathbb{R}^+ \to [-1, \infty)$, $(l, \ell) \mapsto (l - \ell)/\ell$. Thereby, N denotes the fibers' material law for which we make the following assumption.

Assumption 1 *We have that $N \in C^2([-1, \infty), \mathbb{R}_+)$ and for some constant $c \geq -1$ the material law satisfies $N(\varepsilon) = 0$ for $\varepsilon \leq c$ and N is strictly increasing for $\varepsilon > c$.*

This expresses a solely elastic stress-strain behavior, as an increase in stress is associated with further elongation of the fibers. Thereby, c is the strain from which the fibers are under tension. For a material law using $c > 0$, thus, incorporating a zero phase in the stress-strain behavior we refer to [4]. For a strictly increasing choice, implying $c = -1$, we refer to [3] where crimp on the fibers is considered.

2 Graph Structure and Solvability

We consider the model (1)–(2) introduced in [4] and use the representation of the fiber structure as arbitrarily directed graph (e.g., obtained by imposing edge directions according to an underlying node enumeration) to embed it in a compact formulation with characteristic divergence structure. This allows to investigate the solvability and the uniqueness of a solution for the equilibrium model.

Subsequently, $\mathbf{A} \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{E}|}$ denotes the underlying graph's incidence matrix with

$$\mathbf{A}_{i,j} = \begin{cases} -1 & , \text{ if } \ \nu_i = \operatorname{init}(\mu_j) \\ 1 & , \text{ if } \ \nu_i = \operatorname{ter}(\mu_j) \\ 0 & , \text{ else.} \end{cases} \tag{3}$$

Hereby, $\operatorname{init}(\mu_j)$ refers to the start node and $\operatorname{ter}(\mu_j)$ to the end node of edge μ_j. Given an arbitrary node constellation $\mathbf{x} \in \mathbb{R}^{3|\mathcal{N}|}$, the edge vectors can be collectively determined through

$$\mathbf{t}(\mathbf{x}) = \begin{pmatrix} \mathbf{t}_{\mu_1}(\mathbf{x}) \\ \vdots \\ \mathbf{t}_{\mu_{|\mathcal{E}|}}(\mathbf{x}) \end{pmatrix} = (\mathbf{A} \otimes \mathbf{I}_3)^T \mathbf{x} \tag{4}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{t}_\mu(\mathbf{x})$ is the vector representing the directed edge μ and $\mathbf{I}_3 \in \mathbb{R}^{3\times 3}$ is the identity matrix.

In contrast to (2), let $\phi : \mathbb{R}^{3|\mathcal{E}|} \to \mathbb{R}^{3|\mathcal{N}|}$ denote the forces acting in normalized edge direction expressed in terms of the edge vectors collected in $\mathbf{t} \in \mathbb{R}^{3|\mathcal{E}|}$. That is

$$\phi(\mathbf{t}) = \begin{pmatrix} \phi_{\mu_1}(\mathbf{t}) \\ \vdots \\ \phi_{\mu_{|\mathcal{E}|}}(\mathbf{t}) \end{pmatrix}, \text{ with } \phi_\mu(\mathbf{t}) = \frac{\mathbf{t}_\mu}{\|\mathbf{t}_\mu\|_2} N(\epsilon(\, \|\mathbf{t}_\mu\|_2,\ \ell_\mu\,)), \tag{5}$$

where ϕ_μ is continuously continuable in zero for each $\mu \in \mathcal{E}$. To accumulate the forces acting on an interior node $\nu \in \mathcal{N}_I$ according to (2), we add ϕ_μ if μ is an outgoing edge, i.e., $\nu = \operatorname{init}(\mu)$, and subtract it if μ is an incoming edge, i.e., $\nu = \operatorname{ter}(\mu)$. This is to account for the arbitrarily imposed edge directions which yields

$$\sum_{\mu \in \mathcal{E}(\nu)} \mathbf{f}_\mu^\nu(\mathbf{x}) = -\sum_{\mu \in \mathcal{E}} \mathbf{A}_{\nu,\mu} \phi_\mu(\mathbf{t}(\mathbf{x})) = -(\mathbf{A}_{\nu,\cdot} \otimes \mathbf{I}_3)\phi((\mathbf{A} \otimes \mathbf{I}_3)^T \mathbf{x}). \tag{6}$$

Due to (1), the fixation of the boundary nodes, the variables are the positions of the interior nodes only. Let $\mathbf{z} \in \mathbb{R}^{3|\mathcal{N}_I|}$ denote the interior node positions and $\mathbf{g} \in \mathbb{R}^{3|\mathcal{N}_B|}$ that of the boundary nodes. Thus, to express the node positions in terms of $\mathbf{z}$ we introduce

$$\mathbf{x}_\mathbf{g}(\mathbf{z}) = (\mathbf{P}_I \otimes \mathbf{I}_3)^T \mathbf{z} + (\mathbf{P}_B \otimes \mathbf{I}_3)^T \mathbf{g} \tag{7}$$

with orthogonal projections $\mathbf{P}_I \in \mathbb{R}^{|\mathcal{N}_I|\times|\mathcal{N}|}$ and $\mathbf{P}_B \in \mathbb{R}^{|\mathcal{N}_B|\times|\mathcal{N}|}$ onto the interior nodes and boundary nodes, respectively. Then (4) can be expressed in terms of $\mathbf{z}$ through

$$\mathbf{t}(\mathbf{x_g}(\mathbf{z})) = (\mathbf{A}\otimes\mathbf{I}_3)^T\mathbf{x_g}(\mathbf{z}) = (\mathbf{P}_I\mathbf{A}\otimes\mathbf{I}_3)^T\mathbf{z}+(\mathbf{P}_B\mathbf{A}\otimes\mathbf{I}_3)^T\mathbf{g} = \tilde{\mathbf{A}}_I^T\mathbf{z}+\tilde{\mathbf{A}}_B^T\mathbf{g}, \tag{8}$$

where $\tilde{\mathbf{A}}_I = \mathbf{P}_I\mathbf{A} \otimes \mathbf{I}_3$ and $\tilde{\mathbf{A}}_B = \mathbf{P}_B\mathbf{A} \otimes \mathbf{I}_3$ are defined for notational convenience. Apparently, $\tilde{\mathbf{A}}_I \in \mathbb{R}^{3|\mathcal{N}_I|\times 3|\mathcal{E}|}$ and $\tilde{\mathbf{A}}_B \in \mathbb{R}^{3|\mathcal{N}_B|\times 3|\mathcal{E}|}$ are the incidence matrices containing only the rows belonging to interior nodes and boundary nodes, respectively, that are blown up to three dimensions.

Equation (8), also, allows to express (6) in terms of $\mathbf{z}$. Hence, collecting the individual equations (6) for all interior nodes $\nu \in \mathcal{N}_I$ yields the nonlinear system

$$\mathbf{F_g}(\mathbf{z}) := -\tilde{\mathbf{A}}_I\phi(\tilde{\mathbf{A}}_I^T\mathbf{z} + \tilde{\mathbf{A}}_B^T\mathbf{g}) = \mathbf{0}, \tag{9}$$

with $\mathbf{F_g}\colon \mathbb{R}^{3|\mathcal{N}_I|} \to \mathbb{R}^{3|\mathcal{N}_I|}$, which is subsequently referred to as Network Equation System (NES). Each interior node constellation $\mathbf{z}$ satisfying $\mathbf{F_g}(\mathbf{z}) = \mathbf{0}$, for a given boundary node constellation $\mathbf{g}$, meets the conditions (1)–(2).

Particularly noteworthy is the divergence structure in (9), which is similarly found in the context of electrical circuit simulations [2], where the circuit topology determines the solvability of the associated differential-algebraic equations. For the NES, which can be embedded in a quasi-static framework to perform tensile strength simulations, we have the following result.

Theorem 1 *Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be connected and let N satisfy Assumption 1. Then, given a fixed boundary node constellation $\mathbf{g} \in \mathbb{R}^{3|\mathcal{N}_B|}$, we have that*

1. *There exists an interior node constellation $\hat{\mathbf{z}} \in \mathbb{R}^{3|\mathcal{N}_I|}$ with $\mathbf{F_g}(\hat{\mathbf{z}}) = \mathbf{0}$.*
2. *If N is strictly increasing on $[-1, \infty)$ then $\hat{\mathbf{z}} \in \mathbb{R}^{3|\mathcal{N}_I|}$ is an unique solution.*

Proof We show the existence of a potential $E_\mathbf{g}\colon \mathbb{R}^{3|\mathcal{N}_I|} \to \mathbb{R}$, which satisfies $\nabla E_\mathbf{g} = -\mathbf{F_g}$. Then the existence of a minimum to $E_\mathbf{g}$ implies that of a solution to the nonlinear system $\mathbf{F_g}(\mathbf{z}) = \mathbf{0}$ by first order optimality conditions.

For a given constellation of boundary nodes $\mathbf{g} \in \mathbb{R}^{3|\mathcal{N}_B|}$ we define the fiber structure's potential, depending on the interior node positions $\mathbf{z} \in \mathbb{R}^{3|\mathcal{N}_I|}$, through

$$E_\mathbf{g}(\mathbf{z}) = \sum_{\mu\in\mathcal{E}} \ell_\mu G(\varepsilon(\|\mathbf{t}_\mu(\mathbf{x_g}(\mathbf{z}))\|, \ell_\mu)), \quad \text{where} \quad G(\varepsilon) = \int_{-1}^{\varepsilon} N(s)\, ds.$$

That is the weighted sum of the potential energies of the individual fiber connections caused by stretching them. Straightforward application of the chain rule yields

$$
\begin{aligned}
\nabla_{\mathbf{z}} E_{\mathbf{g}}(\mathbf{z}) &= \sum_{\mu\in\mathcal{E}} \ell_\mu \frac{d}{d\varepsilon} G(\varepsilon(\|\mathbf{t}(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))\|, \ell_\mu))\nabla_{\mathbf{z}}\varepsilon(\|\mathbf{t}(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))\|, \ell_\mu))\\
&= \sum_{\mu\in\mathcal{E}} \frac{\mathbf{t}_\mu(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))^T}{\|\mathbf{t}_\mu(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))\|} N(\varepsilon(\|\mathbf{t}(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))\|, \ell_\mu))(\mathbf{A}_{\cdot,\mu}\otimes \mathbf{I}_3)^T(\mathbf{P}_I\otimes \mathbf{I}_3)^T\\
&= \sum_{\mu\in\mathcal{E}} (\mathbf{P}_I\mathbf{A}_{\cdot,\mu}\otimes \mathbf{I}_3)\phi_\mu(\mathbf{t}(\mathbf{x}_{\mathbf{g}}(\mathbf{z})))\\
&= \tilde{\mathbf{A}}_I\phi(\tilde{\mathbf{A}}_I^T\mathbf{z} + \tilde{\mathbf{A}}_B^T\mathbf{g}),
\end{aligned}
$$

which shows that $\mathbf{F}_{\mathbf{g}}$ is the negative gradient field of $E_{\mathbf{g}}$. To verify the existence of a global optimum we show that $E_{\mathbf{g}}$ is coercive, i.e., $E_{\mathbf{g}}(\mathbf{z}) \to \infty$ for $\|\mathbf{z}\| \to \infty$.

Apparently, $\|\mathbf{z}\| \to \infty$ implies $\|\mathbf{x}_\nu\| \to \infty$ for at least one interior node $\nu \in \mathcal{N}_I$. Due to the connectivity of $\mathcal{G}$, we have that any boundary node $\tilde{\nu} \in \mathcal{N}_B$ is connected to ν over a finite path $P = (\mathcal{N}_P, \mathcal{E}_P) \subseteq \mathcal{G}$, with nodes $\mathcal{N}_P = \{\nu_{p_0}, \ldots, \nu_{p_q}\} \subseteq \mathcal{N}$, edges $\mathcal{E}_P = \{(\nu_{p_0}, \nu_{p_1}), \ldots, (\nu_{p_{q-1}}, \nu_{p_q})\} \subseteq \mathcal{E}$ and $q \in \mathbb{N}$ such that $\nu_{p_0} = \nu$ and $\nu_{p_q} = \tilde{\nu}$. As the boundary node $\tilde{\nu} \in \mathcal{N}_B$ is fixed to a given position $\mathbf{g}_{\tilde{\nu}}$, we can conclude

$$
\|\mathbf{x}_\nu - \mathbf{g}_{\tilde{\nu}}\| \leq \sum_{j=1}^{q} \|\mathbf{x}_{\nu_{p_j}} - \mathbf{x}_{\nu_{p_{j-1}}}\| \to \infty, \quad \text{for} \quad \|\mathbf{z}\| \to \infty. \tag{10}
$$

Hence, for at least one $k \in \{1, \ldots, q\}$ it holds that $\|\mathbf{x}_{\nu_{p_k}} - \mathbf{x}_{\nu_{p_{k-1}}}\| \to \infty$, for $\|\mathbf{z}\| \to \infty$, as otherwise we would have a contradiction to (10). Let $\tilde{\mu} = (\nu_{p_k}, \nu_{p_{k-1}})$ denote the respective edge in $\mathcal{E}_P$, then

$$
E_{\mathbf{g}}(\mathbf{z}) \geq \ell_{\tilde{\mu}} G(\varepsilon(\|t_{\tilde{\mu}}(\mathbf{x}_{\mathbf{g}}(\mathbf{z}))\|, \ell_{\tilde{\mu}})) \to \infty, \text{ for } \|\mathbf{z}\| \to \infty, \tag{11}
$$

since Assumption 1 implies $G \geq 0$ and $G(\varepsilon) \to \infty$ for $\varepsilon \to \infty$. Apparently, (11) corresponds to $E_{\mathbf{g}}$ being coercive. Thus, by the continuous differentiability of $E_{\mathbf{g}}$ we can conclude the existence of a global minimum, cf. [1].

Moreover, if N is strictly increasing on $[-1, \infty)$, then $E_{\mathbf{g}}$ is strictly convex, as

$$
\begin{aligned}
G(\varepsilon(\|t_\mu(\mathbf{x}(\lambda\mathbf{z} + (1-\lambda)\tilde{\mathbf{z}}))\|, \ell_\mu)) &= G(\varepsilon(\|\lambda t_\mu(\mathbf{x}(\mathbf{z})) + (1-\lambda)t_\mu(\mathbf{x}(\tilde{\mathbf{z}}))\|, \ell_\mu))\\
&< \lambda G(\varepsilon(\|t_\mu(\mathbf{x}(\mathbf{z}))\|, \ell_\mu)) + (1-\lambda)G(\varepsilon(\|t_\mu(\mathbf{x}(\tilde{\mathbf{z}}))\|, \ell_\mu))
\end{aligned}
$$

for $\lambda \in (0, 1)$ and any pair $\mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^{3|\mathcal{N}_I|}$ with $\mathbf{z} \neq \tilde{\mathbf{z}}$. Here, the equality holds by linearity and the inequality is explained by the fact that ε is convex and that G is strictly increasing and convex. This implies a unique solution for the nonlinear system $\mathbf{F}_{\mathbf{g}}(\hat{\mathbf{z}}) = \mathbf{0}$, cf. [6]. □

3 Structural Analysis

Except for $\mathcal{G}$ being connected, there are no topological restriction to Theorem 1, which is even applicable for multigraphs. This differs from typical requirements for electrical circuit simulation, where additional structural assumptions must be made, e.g., to avoid short circuits. However, we can exploit the topology of the fiber structure to identify subgraphs that have a trivial solution for which the associated equations of the NES are satisfied. This includes loose subgraphs and simple linking nodes, cf. [4], that are subject to following discussion.

Definition 1 A connected subgraph $\mathcal{L} \subset \mathcal{G}$ is refered to as loose if it is connected to the remainder $\mathcal{R} = \mathcal{G} \setminus \mathcal{L}$ over a cutvertex $\nu_c \in \mathcal{N}$ and if it does not contain a boundary node, i.e., $\mathcal{N}(\mathcal{L}) \cap \mathcal{N}_B = \emptyset$.

Loose subgraphs can be neglected, as their constellation is determined by the associated cutvertex. To convince ourselves of this statement, assume that $\mathcal{L}$ is a loose subgraph with associated cutvertex ν_c and remainder $\mathcal{R}$, and that the edges are arranged so that the edges of $\mathcal{R}$ come first. Then we have $\mathbf{A} = [\mathbf{A}_R, \mathbf{A}_L]$, which implies

$$\tilde{\mathbf{A}}_I = [\tilde{\mathbf{A}}_{IR}, \tilde{\mathbf{A}}_{IL}], \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_R \\ \mathbf{t}_L \end{pmatrix}, \quad \text{and} \quad \phi(\mathbf{t}) = \begin{pmatrix} \phi_R(\mathbf{t}_R) \\ \phi_L(\mathbf{t}_L) \end{pmatrix}, \tag{12}$$

where the indices R and L indicate the edges, edge vectors, and acting forces corresponding to the remainder $\mathcal{R}$ and the loose subgraph $\mathcal{L}$, respectively. The information regarding edges connecting the loose subgraphs to the cutvertex is thereby included in the terms indicated by L. Accordingly, $\tilde{\mathbf{A}} = [\tilde{\mathbf{A}}_R, \tilde{\mathbf{A}}_L]$ for $\tilde{\mathbf{A}} = \mathbf{A} \otimes \mathbf{I}_3$. Then, for node constellation $\mathbf{x} \in \mathbb{R}^{3|\mathcal{N}|}$, the NES can be split up, since

$$\tilde{\mathbf{A}}_I \phi(\tilde{\mathbf{A}}^T \mathbf{x}) = \tilde{\mathbf{A}}_{IR} \phi_R(\tilde{\mathbf{A}}_R^T \mathbf{x}) + \tilde{\mathbf{A}}_{IL} \phi_L(\tilde{\mathbf{A}}_L^T \mathbf{x}), \tag{13}$$

where the first term corresponds to the NES associated to the remainder $\mathcal{R}$ and the second term to that of the loose subgraph $\mathcal{L}$. Definition 1 implies that the positions of all nodes in $\mathcal{L}$ are variable and that they are either incident to ν_c or another node in $\mathcal{L}$. Hence, for any $\mathbf{x}$ satisfying $\mathbf{x}_\nu = \mathbf{x}_{\nu_c}$ for all $\nu \in \mathcal{L}$ we have $\mathbf{x} \in \ker(\tilde{\mathbf{A}}_L^T)$ which implies $\phi_L(\tilde{\mathbf{A}}_L^T \mathbf{x}) = \mathbf{0}$ with regard to (5). Hence, for this trivial constellation of loose subgraph nodes the second term in (13) vanishes. Thus, it suffice to determine a solution to the NES of the remainder $\mathcal{R}$, which exists according to Theorem 1.

Definition 2 A node $\nu \in \mathcal{N}_I$ is referred to as simple linking node, if it has degree 2.

Apparently, simple linking nodes link a pair of fiber connections, that can be treated equally as single fiber connection of cumulated length. This can be attributed to the force equilibrium condition (2) and Assumption 1.

For solving the NES, trivial parts of the solution can be neglected, e.g., by removing loose subgraphs and merging fiber connections linked by a simple linking node. Apart from such trivial parts of the solution, it may come to a lack of uniqueness to a solution of the NES when considering a material law that is not strictly increasing. In this case the Newton-Raphson Method may fail, for which a diagonal perturbation of the Jacobian of $\mathbf{F_g}$ can be considered. This corresponds to a Tikhonov Regularization for the minimization of $E_{\mathbf{g}}$, cf. [5]. In the context of nonwoven tensile strength simulations a friction-based regularization approach was introduced in [4] to cope with the ill-posedness of the associated quasi-static simulation approach.

Acknowledgments The support of the German Research Training Group *Algorithmic Optimization – ALOP* (RTG 2126) funded by the DFG is acknowledged.

References

1. Beck, A.: Introduction to Nonlinear Optimization. Society for Industrial and Applied Mathematics, Philadelphia (2014)
2. Estévez Schwarz, D., Tischendorf, C.: Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. (2000)
3. Gramsch, S., Klar, A., Leugering, G., et al.: Aerodynamic web forming: process simulation and material properties. J. Math. Industry (2016)
4. Harmening, M., Marheineke, N., Wegener, R.: Efficient graph-based tensile strength simulations of random fiber structures. Z. Angew. Math. Mech. (2021)
5. Ito, K., Jin, B.: Inverse Problems. World Scientific, Singapore (2014)
6. Ortega, J. M., Rheinboldt, W. C.: Iterative Solution of Nonlinear Equations in Several Variables. Society for Industrial and Applied Mathematics, Philadelphia (2000)

Global-Scale or Fine-Scale Modelling? A Critical Look at Experimental Design

Jochen Wittmann

Abstract Due to the increasing spread of the object-oriented programming paradigm, it is obvious to use a model description on the individual level as a replacement for the traditional differential equation models also for simulation. It promises a system-oriented specification of the model dynamics, especially for users from the application domain. The paper shows that with such an approach, the number of parameters describing the behaviour increases rapidly, which leads to considerable problems in model validation and experimental design. The paper analyses the structure of the problem, lists typical scenarios for model usage, and discusses them concerning parametrisation and validation.

1 Micro and Macro and the Experimental Design

There are different trends concerning the development of modelling and simulation (see some remarks in the overview in [2]), but this paper will focus just on one of them: the availability of large amounts of data as a base for modelling, parameterization, and validation of models. Two main points make the difference to the situation in the past: First, a great amount (and a continuously growing amount) of data is open accessible in the web for everybody. That is surely pushed in general by the open data initiative and especially for the European countries by regulations that demand free access to all the data collected and stored by government agencies. The second point is, that there are really large datasets to exploit for modelling and simulation purposes. The growing technical facilities to store and handle even large datasets opens the access to data of various type, e.g. time series and extensive geodata. So far the good news. From the methodological point of view, these data collections might help to satisfy the demand for experimental data that accompanies every model-based study, but in general all the disposable data sets

J. Wittmann (✉)
Environmental Informatics, University of Applied Sciences (HTW) Berlin, Berlin, Germany
e-mail: wittmann@htw-berlin.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_18

had not been collected with regard to objectives of the model study but more or less accidentally. So we observe growing amount of data, growing dimensionality of the state space the data are collected for and thus the effect, that the data available just points out some islands of information within the multi-dimensional ocean of missing measurements. A more prosaic differentiation can be found in Thiel-Clemen [1]. However, for modelling and simulation pre-planned measurements of the complete state-space with the accuracy determined by the intention of the model are necessary. Thus, the offer of free accessible data on different scales has its disadvantages, too. But not only the data situation leads to multi-scale architectures, but also the trend in modelling methodology itself: There is not only the differential equation approach, but also object-oriented designed models, that mirror the system's structure in the model structure, and even individual-oriented models with their fine-scale approach.(see e.g. [4]) Putting these different approaches together in a common, modular-hierarchical model (like introduced by Zeigler [5] or Eschenbacher [3]), the multi-scale/multi-dimension problems will appear as on the data side: Here the communication and the exchange of data between the model components has to be handled with respect to the changes in scale. This is the problem the following article will focus on. For model development, parameterisation, and validation a change in scale is made to close the gaps arising by missing measurement data on the scale needed originally. The problems on methodological level that are implied by such an experimental design shall be discussed here. To reduce complexity the discussion is made for the situation of a two-scale situation only. The reader might extend the analysis given to the general n-scale problem by simply building all pairwise combinations and handling them as the two-scale one. To understand the problems concerning working with multi-scale models and their uncertainties (such as lack of measurements in some parts of the expanded parameter space, difficulties in measuring the parameters on the fine, additional modelling assumptions), we start with a view on the general design of a modelling and simulation study based on (at least) two model components with different scale that have to be put together into a unified multi-scale model. Both alternatives work according to the same basic scheme: Case A is the situation for a conventional model on global, which means here accumulated, level in specification. The modeler and experimenter are interested in the effects of a change in a global parameter. This parameter is set for the simulation and after the run another parameter on global level, a global indicator variable is observed. On the other hand, case B describes the system dynamics on the fine-scale level. Example: For the population dynamics, a possible input parameter would be the mean number of children a woman gets during her life, one would have to model the interactions of the individuals and would be able to derive an individual curriculum vita for each of the individuals. At the end, the actual number of children each individual has got would be the observation parameter on this level. As long as these levels or scales are only connected by trivial relations such as the summation of the number of children, there are no problems to observe. But any interaction or relation installed between these scales demands for sophisticated treatment as will be shown in the following sections. At this point of the argumentation it should be emphasized that such an interscale-relation does not necessarily have to be an implemented interscale interaction but also might be any connection on argumentative level, e.g. if the data between the scales is compared for validation purposes.

2 Multi-Scale Experimental Design

Problems will arise if the multi-scale approach is not chosen by free decision of the modeler or experimenter but caused by lack of information on the level of data or on the level of model specification. Two cases can be determined: Practical reasons, such as missing data on the scale desired, or even missing knowledge (i.e. missing model) on the scale desired. And secondly, experimental reasons if the results of the simulation are needed for quantities of the other scale or if the research is focused on behavioural aspects of a scale-change in case of emergent behaviour. In all cases, it is necessary to carefully adjust the design of the simulation experiments according to the chosen modelling depth and the available data:

Models on Different Scales If the experimenter works and argues with two separate models on different scale but working independently from each other, there should be no special or additional problems of a multi-scale approach in comparison to the usual approach. Validation and interpretation are just as normal.

Model Components on Different Scales However, if the experimenter works with a modular-hierarchical model, the change of scale is not made after the simulation runs in the phase of analysis and interpretation of the results, but it has continuously be calculated during the simulation run to provide the interface between the different-scale model components connected to each others. For every interaction between the model components involved a scale transformation becomes necessary. Thus, these transformations should be treated with respect to their specification on the value-scale as well as on the time-scale.

Data Transformation Between the Scales There is one observation which appears from the simple description of the experimental set-up described so far: During the simulation run a fine-scale model produces the curriculum vitae of the set of individuals under observation. If the experimenter is interested in more general model quantities, a recalculation and evaluation of those raw data will be necessary. This argumentation implies a change of modelling scale for data evaluation and interpretation (i.e. from level A to level B) concerning the two alternative scenarios introduced in Fig. 1. Similar and more complicated transformations from one level to the other can be necessary in a number of simulation experiments that deal with fine-scaled models. A typical example would be the individual-based approach where fine-scale parameters have to be determined by measured data on global scale; e.g. an individual weight of the model individuals is determined by a measured weight distribution on global scale. In general, the change of scales or

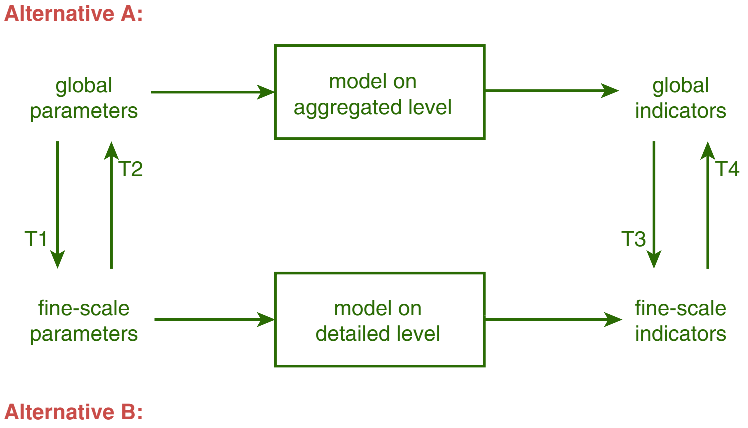


Fig. 1 Possible transformations between the scales during experimentation

levels is successfully applied if missing information on the one scale is replaced by or can be derived from well-known information on the other scale. Such a scale change can be done on the input-side as well as on the side of the outputs. So far there are no problems in the experimental set-up and the situation can be recapitulated graphically by Fig. 1 with transformations T1 to T4. To anticipate the crucial point: The difficulties will arise when the model has to be validated and the situation escalates if there is a lack of comprehensive system data. Usually, the transformations from the individual scale to the global scale are evident and easily to execute. In this direction, there exist data on detail level, which have to be aggregated to a more general, often statistical parameter value on global level. Transformations in the other direction are not possible without at least two further assumptions: First, the type of distribution of the parameter transformed (e.g. uniform, normal, ...), and second the parameters of the distribution, such as mean value, variance, ... But even the very simple transformation of type T4 (individual scale to global scale) might be more than a simple summation and has to be considered with carefulness. An example: The individually collected voices during an election could be weighted. Therefore, an additional set of weight-parameters has to be specified for the model and the corresponding aggregation function has to be calculated for a correctly executed level change. The specification of the scale-change demands more detailed specification. If the scale-change is used because of a lack of information, we see the crux for the experimenter. He has exactly to specify transformation details at a place in the model where there is uncertainty. And naturally, the uncertainty on specification level will cause uncertainty on the level of simulation results. The methodological problem of these parameters is that their values cannot be acquired separately. If it would be possible to do so, the transformation and the scale change would not have been necessary. In the example: If one knows the individual parameters on fine-scale, there would be no need for a change of scale to derive the fine-scale parameters from global scale ones. On the other hand, proper parameter identification needs measurements on both scales to identify the transformation parameters first and to calculate their values afterwards. This is an inherent contradiction of the experimental design. It is caused by the situation of system data and will not be dissolved by additional data acquisition in the real system. Again for the example: The distribution parameters of the global scale can only be known if there are observations on individual scale, too. For the modelling and simulation it follows: A separate validation of the assumptions concerning transformation parameters and their values is not possible. They have to be an additional task within the global model validation process. Thus, the model experiments have to be designed in a manner that the model results are independent of these transformation parameters to have a proper distinction between the influence and effects of the transformations and their parameters and the effects of a change in the model parameters which in fact are under observation. It is obvious that the additional parameters make the study much more complex and the intended direct causality between the experimental parameters and their effects becomes more and more difficult to extract.

3 The Experimental Design Problem for Multi-Scale Models

So far, the need for sophisticated statistical methods for validation has been elaborated. Furthermore, it is obvious that it will not be possible to validate the additional parameters separately, because there are no (or at least: not enough) system data on the desired scales. In this situation, four possible and typical experimental designs shall be analysed with regard on a feasible model validation.

Only Fine-Scale Behaviour Under Observation To be accurate, only the data on fine-scale-level are observed. There is no aggregation of the data at all. Any aggregation would be interpreted as a change to the global scale and would imply the necessity of a transformation of type T4 with the corresponding parameters and difficulties. This leads to the next scenario:

Structural Adequate Models for Global Processes The motivation for this design variant comes from model description methodology: There exists the presumption that a model code as well as a program code is easier to understand and more efficiently to maintain if its structure mirrors the real world structure of the modelled system. If there is a lack of information concerning parameters on the individual level, there are lots of additional hypotheses concerning type and parameter values of the transformations to calculate and validate, a task that has to be solved by data collected on the aggregated scale solely. Thus, a serious validation for this kind of models succeeds only with huge effort in statistical determination of the missing parameters.

Measurements Are Not Possible on the Desired Scale of Model Description This scenario is very similar to the preceding one; however, in this case the experimenter has no choice between the alternatives in scale because a missing access to the data on the one level forces him/her to substitute the missing information by investigations on the other one.

Investigations on Emergent Behaviour It is evident that the use of aggregated scales is useless and the use of fine-scaled models is inevitable in this case. Here, the experiment focuses on one of our transformations: The purpose of the model is to describe individual behaviour on fine-scale, let the individuals interact, and to observe behaviour of the group of individuals that has not been specified explicitly on the local level. The change of level is the trick: input on local, measurement of output on global scale. There is no transformation specification in the form of rules or functions! In contrast, the observations on global level are generated by the behaviour specification on local level exclusively. In real world applications the investigations on emergent behaviour naturally are superposed by the problems in getting proper system data on the scale used for modelling. Therefore, very often level transformations are necessary to avoid data lacks. These transformations have to be parameterised and validated as described before. To prove real evident behaviour properly it is inevitable to separate the effects of the transformation from the observations made to prove the emergent behaviour.

4 Conclusion

The paper tries to give a structure to discuss the problems dealing with uncertainties caused by a multi-scale approach in modelling by mentioning the separate data transformation steps within the global and the local modelling level and between the scales themselves. Of special interest is the discussion, how to use the information available for model-validation purposes. It emphasises that each change of scale causes a transformation with additional parameters for its own that normally have to be determined by additional statistical experiments. If these experiment are not executed, additional uncertainty concerning the values of those additional parameters is brought into the argumentation. A comparison of results gained by models on the different scales may be interesting, however, its statistical value for validation and interpretation of possibly appearing effects is negligible. The proposed scheme does not provide an algorithm to solve the problems in using multi-scaled models but it tries to make the typical structures of argumentation using such models transparent by giving a simple discussion for the two-scale-problem and tries to give a guideline for the discussion of critical aspects and common problems using such types of models. Obviously, the problem demonstrated here with two scales only, has to be widened if a model is composed of more than two different scales. Then the scale-change argumentation has to be applied pairwise to all scale-changes used.

References

1. Thiel-Clemen, Th.: Information Integration in Ecological Informatics and Modelling. In: Wittmann, J.; Müller, M.(eds.) Simulation in Umwelt- und Geowissenschaften, Shaker, Aachen (2013), pp. 89–96
2. Wittmann, J.: Environmental Modeling and Simulation: A subjective update of the state of the art. In: Pillmann, W. et al. (eds.) Innovations in Sharing Environmental Observations and Information, Shaker, Aachen (2011), pp. 453–459
3. Eschenbacher, P.: Entwurf und Implementierung einer formalen Sprache zur Beschreibung dynamischer Modelle, Dissertation, Universität Erlangen (1990)
4. Ortmann, J.: Ein allgemeiner individuenorientierter Ansatz zur Modellierung von Populationsdynamiken in Ökosystemen unter Einbeziehung der Mikro- und Makroebene, Dissertation am Fachbereich Informatik, Universität Rostock (1999)
5. Zeigler, B.P.: Object-oriented Simulation with Hierarchical, Modular Models, Academic Press, London, (1990)

An Isogeometric One-Dimensional Model for Developable Flexible Elastic Strips

Benjamin Bauer, Michael Roller, Joachim Linn, and Bernd Simeon

Abstract This paper aims at introducing a kinematical reduction for Kirchhoff-Love shells with developable base surfaces that undergo isometric deformations. This framework is appropriate to model, for example, flexible flat cables. In order to decrease the involved number of degrees of freedom, we utilise kinematical reduction to a geodesic line and a vector field along this curve. Application of a relatively parallel frame allows us to generalise this framework to a more general class of curves that may exhibit points or segments of vanishing curvature. We derive the one-dimensional bending energy functional for a rectangular strip, combine it with penalty terms addressing the nonlinear constraints, and compute the equilibrium state as minimiser of this penalised energy. An isogeometric discretisation yields finitely many degrees of freedom for the inner point optimiser. Several example strips clamped at both ends illustrate the feasibility of this approach.

1 Introduction

Thin sheet-like components belong to the most frequently used structural parts in engineering applications. For example, flexible flat cables are of considerable interest in the development of consumer electronics. In order to model their high flexibility and elastic behaviour, the digitalisation of industrial processes relies on physically correct models and efficient numerical methods.

Classical shell theories [1] model thin-walled objects based on their centre surface, thereby reducing both number of involved degrees of freedom and numer-

B. Bauer (✉) · M. Roller · J. Linn
Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
e-mail: benjamin.bauer@itwm.fraunhofer.de; michael.roller@itwm.fraunhofer.de; joachim.linn@itwm.fraunhofer.de

B. Simeon
Felix-Klein Zentrum, TU Kaiserslautern, Kaiserslautern, Germany
e-mail: simeon@mathematik.uni-kl.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_19

ical costs. Fosdick and Fried [4] collected approaches to continue this idea of dimensional reduction for base surfaces, which can be flattened to the plane without change of metric: so-called developable surfaces. Sadowsky [10] and Wunderlich [14] considered the analytic integration of the bending energy of such an infinitesimally narrow strip along its width dimension. Starostin and van der Heijden [11] recently proposed a one-dimensional model for bands. They represent the base surface of a Kirchhoff-Love shell by a rectifying developable (RD) of its centre curve.

In Sect. 2, we avoid the strict requirements of a Frenet frame for this approach and generalise the concept of RDs to curves which may inhibit singularities in the form of vanishing curvature. A relatively parallel frame [2] (in literature also called rotation minimising frame, parallel transport frame or Bishop frame) allows us to decompose the director of the developable and thereby ruled surface. As we consider isometric deformations of the centre surface, the stored energy function consists only of the bending energy. We analytically integrate this energy along the width dimension and end up with a result similar to [14] in Sect. 3. An interior point optimiser [13] then computes the static equilibrium state under geometric boundary conditions where the highly non-linear constraints are addressed by a penalty method. Section 4 describes these numerical details and we display and discuss our results in Sect. 5.

2 Generalised Rectifying Developable Surfaces

Every developable surface is ruled, that means for length L it can be represented by a regular base curve $\boldsymbol{\gamma}\colon [0, L] \to \mathbb{R}^3$ parametrised by arc-length and a director vector field $\mathbf{d}\colon [0, L] \to \mathbb{R}^3$ in the form

$$\boldsymbol{\phi}\colon Q \to \mathbb{R}^3, \qquad (s, v) \mapsto \boldsymbol{\gamma}(s) + v\mathbf{d}(s) \tag{1}$$

with parameters in

$$Q = \{(s, v) \mid 0 \le s \le L,\ v_1(s) \le v \le v_2(s)\}.$$

Note that, in general, the interval bounds v_1, v_2 for v are allowed to vary with respect to the first parameter s.

A ruled surface is developable if and only if the determinant formed by curve tangent $\mathbf{t}$, director and director derivative $\det[\mathbf{t}, \mathbf{d}, \mathbf{d}']$ vanishes everywhere along the curve [12, chap. 5.5]. Following the notation of [12], we indicate derivatives with respect to the arc-length parameter s of $\boldsymbol{\gamma}$ with a prime as in $\mathbf{d}'$ and general derivatives with a dot as in $\dot{\mathbf{d}}$.

The rectifying developable surface (RD) of a curve is the envelope of rectifying planes, i.e. those planes spanned by the curve tangent and the Frenet binormal. By this construction, the curve is a geodesic on its RD and the surface is developable

[12]. Paired with linear independence of director and curve tangent, these properties are characteristic for the RD [6, Proposition 4.1]. However, this framework requires a well-defined Frenet frame. For generalisation let $\mathcal{C}^k(X; Y)$ denote the space of k-times continuously differentiable functions from X to Y.

Definition 1 Let $\boldsymbol{\gamma} \in \mathcal{C}^2([0, L]; \mathbb{R}^3)$ be a curve and $\mathbf{d} \in \mathcal{C}^1([0, L]; \mathbb{R}^3)$ a vector field. Then we refer to the ruled surface $\boldsymbol{\phi}$ constructed by (1) as *generalised rectifying developable* (GRD) if all subsequent conditions are fulfilled:

(a) $\boldsymbol{\phi}$ is developable,
(b) $\boldsymbol{\gamma}$ is a geodesic on $\boldsymbol{\phi}$ and
(c) curve tangent $\mathbf{t}$ and director $\mathbf{d}$ are pointwise linear independent.

Condition (b) is equivalent to a vanishing geodesic curvature κ_g of $\boldsymbol{\gamma}$ within $\boldsymbol{\phi}$ [3]. From now on, we will always assume $\boldsymbol{\gamma}$ and $\mathbf{d}$ to be of the smoothness required by Definition 1.

Note that the three conditions allow the director to scale arbitrarily. In order to parametrise a rectangular strip with non-varying width $2w$, the bounds v_1, v_2 then need to adapt to the director length. However, they may be chosen constant as $-v_1 \equiv w \equiv v_2$ if and only if the projection of the director to the normal plane is of unit length, i.e. $\|\mathbf{d} - (\mathbf{t} \cdot \mathbf{d})\mathbf{t}\| \equiv 1$. The sign $\equiv$ denotes pointwise equality for all $s \in [0, L]$. We end up with a system of equations representing the requirements for the GRD:

$$\det[\mathbf{t}, \mathbf{d}, \mathbf{d}'] \equiv 0, \qquad \kappa_g \equiv 0, \qquad \|\mathbf{t} \times \mathbf{d}\| \equiv 1. \tag{2}$$

Additionally, we pay attention to singularities induced by intersection points of rulings. These form the so-called edge of regression (cf. [12, chap. 5.1]). Let the director be decomposed along a relatively parallel frame $(\mathbf{t}, \mathbf{m}_1, \mathbf{m}_2)$ [2] as

$$\mathbf{d} \equiv d_0\mathbf{t} + d_1\mathbf{m}_1 + d_2\mathbf{m}_2. \tag{3}$$

Such a frame exists even for curves with slightly lesser smoothness requirements [7]. The strip width is bounded by the minimum value of $|d_0'|^{-1}$ and, vice versa, the absolute value of d_0' needs to be bounded by $\frac{1}{w}$. With the decomposition (3) of the director, the requirements for the GRD (2) may be formulated in terms of the director coordinate functions d_i and the curvature components k_1, k_2 with respect to the relatively parallel frame:

$$\det[\mathbf{t}, \mathbf{d}, \mathbf{d}'] \equiv d_1\left(d_2' + d_0k_2\right) - d_2\left(d_1' + d_0k_1\right) \equiv 0 \tag{4a}$$

$$\kappa_g \equiv d_1k_1 + d_2k_2 \equiv 0 \tag{4b}$$

$$\|\mathbf{d} - (\mathbf{t} \cdot \mathbf{d})\mathbf{t}\|^2 \equiv d_1^2 + d_2^2 \equiv 1 \tag{4c}$$

3 The Elastic Bending Energy

Consider a rectangular GRD $\boldsymbol{\phi}$ of its centreline $\boldsymbol{\gamma}$ parametrised by (1) where the director fulfils (2). Since we consider isometric deformations of the centre surface, there is no membrane energy involved and the stored energy functional consists only of the shell bending energy. We mimic the steps of [11, Sect. 3] and derive the one-dimensional energy functional for homogeneous isotropic material and plane stress-free reference configuration

$$\begin{aligned}\Xi = \frac{D}{2}\iint_{\boldsymbol{\phi}} H^2 \mathrm{d}A &= \frac{D}{2}\int_0^L (d_1k_2 - d_2k_1)^2\left(d_0^2+1\right)^2 \int_{-w}^{w} \frac{1}{1+vd_0'}\,\mathrm{d}v\,\mathrm{d}s \\ &= Dw\int_0^L (d_1k_2 - d_2k_1)^2\left(d_0^2+1\right)^2\, V(wd_0')\,\mathrm{d}s.\end{aligned} \tag{5}$$

Here, D denotes the flexural rigidity of the isotropic material, H the mean curvature and

$$V(wd_0') \equiv \frac{1}{wd_0'}\log\left(\frac{1+wd_0'}{1-wd_0'}\right) \equiv 1 + \mathcal{O}((wd_0')^2)$$

gives the small width approximation term which may be neglected under linearisation about an infinitely narrow band [14].

4 The Numerical Model

Based on the implicit formulation for GRDs from Sect. 2 we minimise the energy functional (5) regarding the constraints (4) and the regularity condition $\left|d_0'\right| < \frac{1}{w}$ everywhere. The base curve $\boldsymbol{\gamma}$ and the director field $\mathbf{d}$ constitute the degrees of freedom, where latter is represented by the coefficient functions d_i.

The condition (4c) yields the existence of an angle σ such that $d_1 \equiv -\sin\sigma$ and $d_2 \equiv \cos\sigma$. Geometrically, σ denotes the angle between relatively parallel and Frenet frame wherever latter exists.

In order to solve the underlying optimisation problem, we discretise $\boldsymbol{\gamma}$, d_0 and σ by isogeometric curves of identical basis functions. Based on the number of control points $n+1$ and the degree p chosen by the user, a clamped knot vector $U = [u_0, \ldots, u_{n+p+1}]$ with

$$0 = u_0 = \ldots u_p < u_{p+1} \leq \cdots \leq u_n < u_{n+1} = \cdots = u_{n+p+1} = L$$

defines the family of B-Spline basis functions $\{N_{i,p}\colon [0,L] \to [0,1] \mid i = 0,\ldots,n\}$. For details on the construction and properties confer [9]. A B-Spline curve $\boldsymbol{\beta}$ in q dimensions then reads

$$\boldsymbol{\beta}\colon [0, L] \to \mathbb{R}^q,\ u \mapsto \sum_{i=0}^{n} N_{i,p}(u)\mathbf{p}_i,$$

where $\mathbf{p}_i \in \mathbb{R}^q$ are the control points, which become the degrees of freedom for our approach. The construction of rational B-Spline curves, which we employ for the discretisation of $\boldsymbol{\gamma}$, may again be looked up in [9].

A penalty approach addresses the highly nonlinear constraints (4a),(4b) and prevention of tensile stress:

$$\min_{\boldsymbol{\gamma}, d_0, \sigma} \quad \Xi + \lambda_1\Psi + \lambda_2\Theta + \lambda_3\Omega,$$

$$\Psi = \int_0^L \det[\mathbf{t}, \mathbf{d}, \mathbf{d}']^2 \mathrm{d}s, \qquad \Theta = \int_0^L \kappa_g^2 \mathrm{d}s, \qquad \Omega = \int_0^L \left(\|\dot{\boldsymbol{\gamma}}\| - 1\right)^2 \mathrm{d}u.$$

In this formulation, $\lambda_1 \to \infty$ ensures developability, $\lambda_2 \to \infty$ maintains the geodesic property of $\boldsymbol{\gamma}$ and $\lambda_3 \to \infty$ enforces the inextensibility of the centre curve.

5 Results and Conclusion

We test the model described in the previous section on several benchmarks. Each time, the input consists of a flat reference geometry positioned such that length and width align with x- and y-axis, respectively. Furthermore, we specify geometric boundary conditions in the end points.

The automatic differentiation library autodiff [8] provides gradients and Hessian matrices. The interested reader may confer [5] for theoretical details on this method. The interior point optimiser IPopt [13] takes on the minimisation with respect to boundary conditions.

Figure 1 illustrates an example where position, tangent and director in the first end point ($u = 0$) remain fixed. Then the second end point (in $u = L$) is dragged both upwards in z-direction and towards the first end point in x-direction, leaving

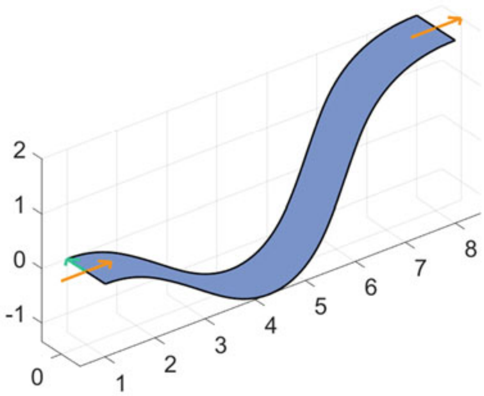


Fig. 1 Equilibrium configuration under symmetric boundary conditions

the tangent fixed but the director free. Under these boundary conditions, the problem reduces to two dimensions where the director is transported parallelly along the plane curve. Note that the method does not struggle with the equivalent buckling state resulting from point reflection and automatically converges to the equilibrium closer to the initial state.

For the second example, we switch both tangents upwards and drag the second end point close to the first. As before, the boundary conditions allow for a planar generator curve. The resulting circular shape is depicted in Fig. 2. This use case demonstrates that the model may cope with large deformations without having the user to specify intermediate steps.

The boundary conditions of the third use case enforce a spatial generator curve. Additionally to the movements of the first example, we translate the second end in y-direction and let the corresponding tangent point in this direction, as well. Figure 3 illustrates the resulting equilibrium state.

Although all examples yield suiting equilibrium states, the model proves itself to be ill-conditioned and susceptible to slow convergence. The condition numbers of the Hessian matrix within the last iteration for the three examples read $1.8 \cdot 10^{20}$, $1.2 \cdot 10^{17}$, $2.1 \cdot 10^{17}$ respectively. Thus, real time applications require future work on the numerics in order to speed up the solution process. Furthermore, gradients may theoretically be derived by solving an ordinary differential equation related to the relatively parallel transport of normal vectors in order to abstain from automatic differentiation.

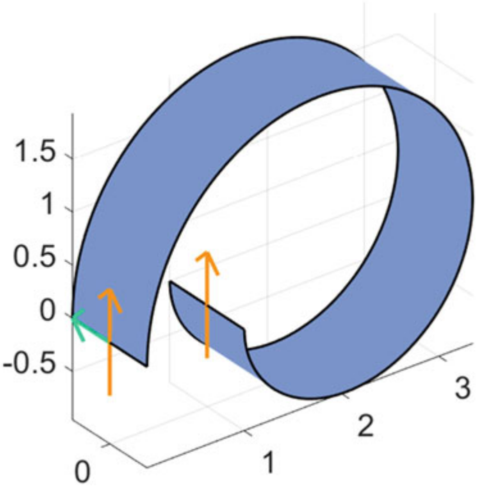


Fig. 2 Circular equilibrium configuration after large deformation

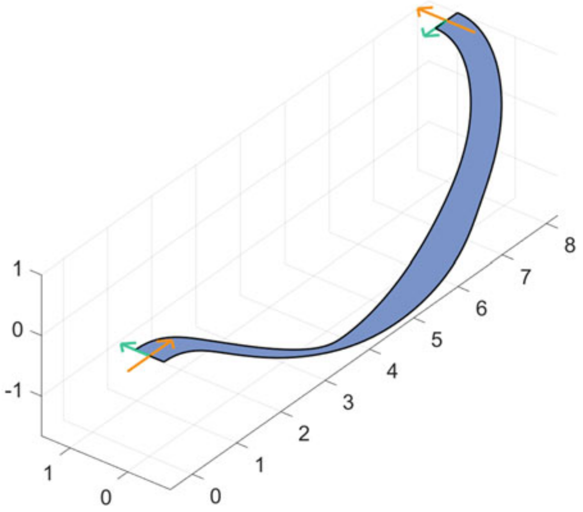


Fig. 3 Example with non-planar centre curve

References

1. Bischoff, M., Wall, W.A., Bletzinger, K.-U., Ramm, E.: Models and Finite Elements for Thin-walled Structures. Encycl. Comput. Mech., **172**, 59–137 (2004)
2. Bishop, R.L.: There is More than One Way to Frame a Curve. Am. Math. Mon. **82**(3), 246–251 (1975)
3. do Carmo, M.P.: Differential geometry of curves and surfaces. Prentice-Hall, Englewood Cliffs, New Jersey (1976)
4. Fosdick, R., Fried, E.: The Mechanics of Ribbons and Möbius Bands. Springer, Netherlands (2016)
5. Griewank, A., Walther, A.: Evaluating Derivatives, Principles and Techniques of Algorithmic Differentiation. SIAM (2008)
6. Izumiya, S., Takeuchi, N.: New Special Curves and Developable Surfaces. Turk. J. Math. **28**, 153–163 (2004)
7. Krejčiřík, D., Šediváková, H.: The Effective Hamiltonian in Curved Quantum Waveguides under mild Regularity Assumptions. Rev. Math. Phys. **24** (2012)
8. Leal, A.M.M.: Autodiff, a modern, fast and expressive C++ library for automatic differentiation. https://autodiff.github.io (2018)
9. Piegl, L., Tiller, W.: The NURBS Book. Springer, Berlin, Heidelberg (1995)
10. Sadowsky, M.: Ein elementarer Beweis für die Existenz eines abwickelbaren Möbiusschen Bandes und Zurückführung des geometrischen Problems auf ein Variationsproblem. Sitzungsberichte der Preussischen Akademie der Wissenschaften. Philos.-hist. Klasse (1930)
11. Starostin, E.L., van der Heijden, G.H.M.: Equilibrium Shapes with Stress Localisation for Inextensible Elastic Möbius and Other Strips. J. Elast. **119**(1), 67–112 (2015)
12. Struik, D.J.: Lectures on classical differential geometry. Dover Publications (1988)
13. Wächter, A., Biegler, L.T.: On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming. Math. Program., **106**(1), pp. 25-57 (2006)
14. Wunderlich, W.: Über ein abwickelbares Möbiusband. Monatsh. Math. **66**(3), 276–289 (1962)

A Hybrid DEIM and Leverage Scores Based Method for CUR Index Selection

Perfect Y. Gidisu and Michiel E. Hochstenbach

Abstract The discrete empirical interpolation method (DEIM) may be used as an index selection strategy for formulating a CUR factorization. A notable drawback of the original DEIM algorithm is that the number of column or row indices that can be selected is limited to the number of input singular vectors. We propose a new variant of DEIM, which we call L-DEIM, a combination of the strength of deterministic leverage scores and DEIM. This method allows for the selection of a number of indices greater than the number of input singular vectors. Since DEIM requires singular vectors as input matrices, L-DEIM is particularly attractive for example in big data problems when computing a rank-k SVD approximation is expensive even for moderately small k since it uses a lower-rank SVD approximation instead of the full rank-k SVD. We empirically demonstrate the performance of L-DEIM, which despite its efficiency, may achieve comparable results to the original DEIM and even better approximations than some state-of-the-art methods.

1 Introduction

Data sets are often represented by large matrices. In recent times, with the growth of the internet (industrial) data matrices are big and may be hard to manage. Examples of such data sets include text documents, customer databases, stocks, and financial transactions. In many data analyses, we need dimension reduction and for many applications, we need interpretable dimension reduction of which a CUR decomposition is one form. A CUR factorization is a low-rank matrix approximation proposed as an alternative to the TSVD to ensure interpretability and preserve relevant properties like sparsity or nonnegativity of the underlying matrix. A rank-k CUR decomposition of an $m \times n$ matrix A has the form

P. Y. Gidisu (✉) · M. E. Hochstenbach
TU Eindhoven, Eindhoven, The Netherlands
e-mail: p.gidisu@tue.nl; m.e.hochstenbach@tue.nl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_20

$$A \approx CMR = AP \cdot M \cdot S^T A,$$

where $C \in \mathbb{R}^{m\times k}$ and $R \in \mathbb{R}^{k\times n}$ are subsets of the columns and rows of A, respectively. The matrices $P \in \mathbb{R}^{n\times k}$ and $S \in \mathbb{R}^{m\times k}$ are index selection matrices with some columns of the identity indicating the columns and rows that are picked. The matrix M is constructed to minimize the approximation error. There are several variants of this decomposition, which implies that the three factors are not necessarily unique. In [7, 9] the authors present algorithms for a CUR factorization based on a rank-k singular value decomposition. Sorensen and Embree [9] propose a CUR approximation using a discrete interpolation method (DEIM) on the rank-k singular vectors. The index selection method DEIM has first been introduced in the context of model order reduction [1]. In [9], it is shown to be a viable index selection method for identifying the most representative and influential subset of columns and rows that define a low-dimensional space of the data. The DEIM-induced CUR requires the computation of the SVD or its approximation. A notable limitation of this index selection algorithm is that the number of indices that can be selected is limited to the number of available singular vectors. In an attempt to address this, we propose a new extension called L-DEIM. The L-DEIM scheme combines the strengths of deterministic leverage scores sampling [8] and the DEIM procedure. Our new approach is an alternative index selection method that is particularly attractive in a setting (for example big data problems) where we want a rank-k CUR decomposition and computing a rank-k SVD approximation is expensive even for moderately small k. This new algorithm allows us to select k indices without having to compute the full rank-k SVD by using a lower-rank SVD approximation instead. It may be viewed as an approach to reuse the same information to further improve the approximation.

We denote the 2-norm by $\|\cdot\|$. We use MATLAB notation to index vectors and matrices; thus, $A(:, p)$ denotes the k columns of A whose corresponding indices are in vector $p \in \mathbb{N}_+^k$.

2 Related Works

In this section, we briefly review some state-of-the-art deterministic algorithms for a CUR decomposition. These algorithms have been developed for the column subset selection problem or interpolative decomposition, but can be generalized for a CUR decomposition. We derive our proposed algorithm L-DEIM by combining two of the algorithms.

2.1 Standard DEIM

The DEIM is a discrete variant of the empirical interpolation method for approximating systems of nonlinear ordinary differential equations. In a recent paper by Sorensen and Embree [9], the authors use this method in the formulation of a CUR decomposition. The DEIM algorithm requires a full rank-k SVD of A to select at most k column and or row indices of A. To illustrate how the indices are selected via the DEIM index selection method, we first define a projector which the authors in [1] called an interpolatory projector. Suppose we want to preserve k rows of A and we have the rank-k approximation of A as

$$\underset{m \times n}{A} \approx \underset{m \times k}{U} \; \underset{k \times n}{F},$$

where U contains the top k left singular vectors. The matrix F is a coefficient matrix to be defined such that the above approximation preserves exactly the desired k rows of A. Let $\mathbf{s} \in \mathbb{N}_+^k$ be an index vector with unique entries from the row index set $\{1, \ldots, m\}$ of A. Now let $S \in \mathbb{R}^{m \times k}$ be an index selection matrix with some columns of the identity matrix that selects certain rows of A, i.e., $S = I(:, \mathbf{s})$. Assuming we want to keep desired rows in $\mathbf{s}$ in the approximation, viz., $S^T A \approx S^T(UF)$. If $S^T U$ is nonsingular, the coefficient matrix F can be determined uniquely; $F = (S^T U)^{-1} S^T A$. This implies $A \approx U(S^T U)^{-1} S^T A = \mathbb{S}A$. The operator $\mathbb{S}$ is the DEIM interpolatory projector, an oblique projector. The name interpolatory comes from the fact that the projected matrix $\mathbb{S}A$ matches A in the $\mathbf{s}$ entries. Note that we can obtain a similar projector using the right singular vectors. The DEIM algorithm processes the left singular vectors sequentially starting with the first dominant singular vector. Each step considers the next singular vector to obtain the next index. The selected indices are used to compute the interpolatory projector $\mathbb{S}$. The next index is selected by removing the direction of the interpolatory projection in the previous vectors from the subsequent one and finding the index of the entry with the largest magnitude in the residual vector (for more details see [9]).

In [2], Drmac and Gugercin proposed the Q-DEIM; a variant of DEIM which runs a column pivoted QR factorization on the transposes of the right and left singular vectors to select the column and row indices, respectively.

2.2 *Deterministic Leverage Score Sampling*

Part of the new extension borrows an idea from the leverage scores of a matrix A, which is defined below. We denote the ith row of V_k by $[V_k]_{i,:}$.

Definition 1 Given a matrix $A \in \mathbb{R}^{m \times n}$ with rank$(A) \geq k$, let V_k contain its k leading right singular vectors. The rank-k leverage score of the ith column of A is

$$\ell_i = \|[V_k]_{i,:}\|^2, \quad i = 1, \ldots, n.$$

The deterministic leverage score sampling procedure selects columns of A corresponding to the indices of the largest leverage scores for a given k. This deterministic column selection method proposed by Jolliffe [6] is one of the first column subset selection algorithms. The leverage score sampling algorithm can extract at least k column indices of A and the upper bound on the number of indices that can be selected is not immediate. For more details on the algorithm and the bound on the number of columns to be sampled see, [8, Sect. 3.1].

3 L-DEIM

We now introduce the new extension of DEIM. Our starting point is the method from the earlier work [9], which derives a rank-$\widehat{k}$ CUR factorization by applying DEIM to the $\widehat{k}$ singular vectors. Given the promising results of this algorithm compared to other state-of-the-art methods for a CUR approximation, our proposed algorithm builds on the DEIM procedure. Constructing a rank-$\widehat{k}$ CUR decomposition using L-DEIM requires a rank-k singular vectors where $\widehat{k} > k$. The integer k is the number of available (approximate) singular vectors, while $\widehat{k}$ is the number of indices to be selected. To select the $\widehat{k}$ indices, the proposed method performs the original DEIM to find the first k indices while keeping the residual singular vector in each index selection step of the DEIM procedure. The residual singular vector is the error between the input singular vector and its approximation from interpolating the previous singular vectors at the selected indices; as in line 2 of Algorithm 1.[1] At the end of the iteration, using the idea of leverage scores, we compute the 2-norm of the rows of the residual singular vectors to select the additional $\widehat{k} - k$ indices. The procedure is summarized in Algorithm 1. Note that the vectors in U in line 3 of Algorithm 1 are the residual singular vectors and not the original singular vectors.

Algorithm 1: L-DEIM index selection

```
Input: $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$, target rank = $\widehat{k}$, with $k \leq \widehat{k} \leq \min(m, n)$
Output: column and row indices $\mathbf{s}, \mathbf{p} \in \mathbb{N}_+^k$, respectively, with non-repeating entries
 for $j = 1, \dots, k$
1 :  $\mathbf{s}(j) = \operatorname{argmax}_{1 \leq i \leq m} |(U(:, j))_i|$
2 :  if $j < k$; $U(:, j+1) = U(:, j+1) - U(:, 1:j) \cdot (U(\mathbf{s}, 1:j) \setminus U(\mathbf{s}, j+1))$
3 : Compute $\ell_i = \|[U]_{i:}\|$   for $i = 1, \dots, m$;  sort $\ell$ in non-increasing order
4 : Remove entries in $\ell$ corresponding to the indices in $\mathbf{s}$
5 : $\mathbf{s}' = \widehat{k} - k$ indices corresponding to $\widehat{k} - k$ largest entries of $\ell$
6 : $\mathbf{s} = [\mathbf{s};\ \mathbf{s}']$
7 : Perform 1–6 on $V$ to get index set $\mathbf{p}$
```

[1] Note that the backslash operator used in the algorithm is a Matlab type notation for solving linear systems and least-squares problems.

From Algorithm 1, if $\widehat{k} = k$ then the algorithm reduces to the standard DEIM. We note that if the target rank is not specified, given k, we can select at least k indices but the upper bound on the number of indices to be selected is not immediate; we can select an arbitrary number of indices. Similar to leverage scores sampling, the L-DEIM allows for oversampling of columns and or rows.

Error Bounds Let us consider a fixed matrix $A \in \mathbb{R}^{m\times n}$ with rank $\rho \leq \min(m, n)$. For an arbitrary k with $1 \leq k \leq \rho$, the best rank-k approximation of A (A_k) provided by the SVD gives $\|A - A_k\| = \sigma_{k+1}(A)$ where σ_{k+1} is the $(k + 1)$st singular value of A. Suppose that we have a known target rank $k < \min(m, n)$, a good rank-$\widehat{k}$ approximation $A_{\widehat{k}}$ gives $\|A - A_{\widehat{k}}\| \leq \tau \; \|A - A_k\|$, where $\tau > 0$ is a modest tolerance and $k \leq \widehat{k} \leq r$ is the rank of the decomposition with oversampling. The following result unifies the theoretical bound results for $\|A - CMR\|$ in [9, Sect. 4] and [5, Append. 1].

Proposition 1 (See [9, Sect. 4], [5, Append. 1]) *Given $A \in \mathbb{R}^{m\times n}$ and $1 \leq k \leq \widehat{k} \leq \min(m, n)$, let $S \in \mathbb{R}^{m\times\widehat{k}}$, $P \in \mathbb{R}^{n\times\widehat{k}}$ be index selection matrices and the top k left and $right$ singular vectors be $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$, respectively. Let $C = AP \in \mathbb{R}^{m\times\widehat{k}}$ and $R = S^T A \in \mathbb{R}^{n\times\widehat{k}}$ be of full rank, assuming we compute M as $(C^T C)^{-1}C^T A\; R^T(RR^T)^{-1}$ and $S^T U$ and $V^T P$ are of full rank we have*

$$\|A - CMR\| = [\sigma_{\min}^{-1}(V^T P) + \sigma_{\min}^{-1}(S^T U)] \cdot \sigma_{k+1}.$$

The above error bounds suggest an index selection method which minimizes the quantities $\sigma_{\min}^{-1}\,(V^T P)$ and $\sigma_{\min}^{-1}\,(S^T U)$ is theoretically desirable.

4 Experiments

We perform some experiments to compare the approximation quality and runtimes of the new method L-DEIM with the existing deterministic methods discussed in Sect. 2. We use the relative error $\|A - CMR\|/\|A\|$ and runtimes for selecting the column and row indices as the evaluation criteria. Note that the runtimes reported here do not include the time for computing the singular vectors. We run the algorithms on three real data sets used in [7, 9]. The application domains of the data sets are Internet term document analysis, genetics, and collaborative filtering. The Internet term document data is from the Technion Repository of Text Categorization Datasets (TechTC). We use test 26, which consists of a collection of 139 documents on two topics with 15,210 terms describing each document [3]. As in [9], the 139 × 15,210 TechTC matrix rows are scaled to have a unit 2-norm. We take the cancer genetics data set GSE10072 from National Institutes of Health. This data set has 107 patients described by 22,283 probes. There are 58 patients with tumors and 49 without. We center the 22,283 × 107 genetics data matrix by subtracting the mean of each row from the entries in that row. The final data set is the Jester joke data set

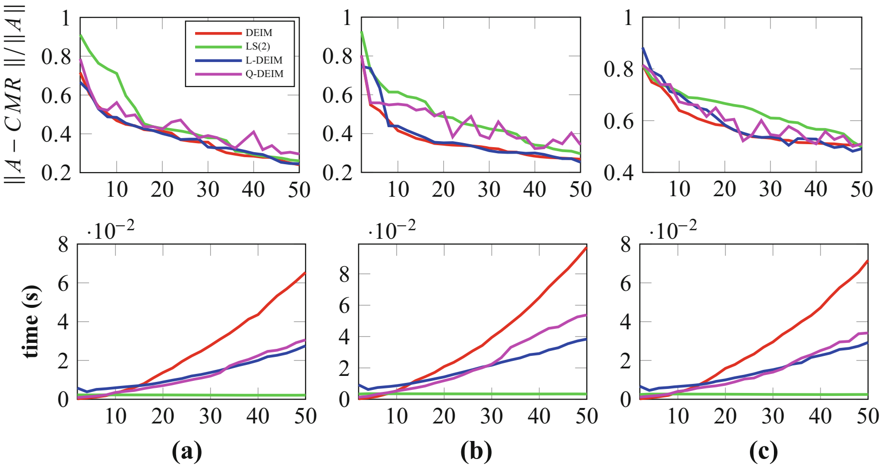


Fig. 1 The approximation quality (first row) and runtimes (second row) of the L-DEIM scheme compared with the standard DEIM, Q-DEIM, and leverage scores sampling techniques using the three real data sets. Displayed are the relative errors $\|A - CMR\|/\|A\|$ and runtimes as a function of rank k. (**a**) Jester jokes data. (**b**) Cancer genetics data. (**c**) TechTC text data

[4], which is often used as a benchmark for recommender system research. The data matrix consists of 73,421 users and their ratings for 100 jokes. We only consider users who have ratings for all 100 jokes. We center the resulting 14,116 × 100 matrix by subtracting the mean of each column from all entries in that column.

From Fig. 1, we see that the approximation quality of the proposed method L-DEIM is as good as the original DEIM while the L-DEIM enjoys favorable runtimes. Both DEIM and L-DEIM have considerably lower approximation error than the other methods. The leverage scores sampling using two singular vectors seems to be the most efficient; however, we note that there is a trade-off between the runtimes and approximation quality. We show results of the leverage scores method using only the leading two singular vectors since higher choices yield worse approximation results.

5 Conclusions

We have presented a new extension of the DEIM index selection algorithm (L-DEIM) to identify additional indices for constructing a rank-$\widehat{k}$ CUR decomposition using a lower-rank SVD approximation. This is especially useful in a setting (for example big data problems) where computing a full rank-$\widehat{k}$ SVD is relatively expensive. The algorithm may be viewed not only as an extension of DEIM but also as an alternative index selection method for a CUR factorization. The L-DEIM procedure may also be suitable for point selection in the context of model order for nonlinear dynamical systems. Although the proposed algorithm is computationally more efficient than the original DEIM, experiments show that the approximation accuracy of both methods may be comparable when the target rank $\widehat{k}$ is at most twice the available k singular vectors. For all results presented in Sect. 4, we assume that given a target rank $\widehat{k}$, $2k = \widehat{k}$ in Algorithm 1. From experiments not presented here, if $\widehat{k} > 2k$ in Algorithm 1, then the rank-$\widehat{k}$ CUR approximation quality of the L-DEIM procedure which uses k singular vectors may generally be worse than the rank-$\widehat{k}$ CUR factorization quality of the standard DEIM scheme which requires $\widehat{k}$ singular vectors. However, we stress that the L-DEIM is considerably cheaper. A code for L-DEIM is available on https://github.com/perfectyayra/L-DEIM-index-selection.

Acknowledgments This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812912.

References

1. S. Chaturantabut and D. C. Sorensen: *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput. 32 (2010), pp. 2737–2764.
2. Z. Drmac and S. Gugercin: *A New selection operator for the discrete empirical interpolation method-improved a priori error bound and extensions*, SIAM J. Sci. Comput. 38 (2016), pp. A631–A648.
3. E. Gabrilovich and S. Markovitch (2004) Data from: Technion Repository of Text Categorization Datasets. http://gabrilovich.com/resources/data/techtc/
4. K. Goldberg et al. (2001) Data from: UC Berkeley AutoLab. http://eigentaste.berkeley.edu/dataset/
5. E.P. Hendryx, B.M. Rivière and C.G. Rusin: *An extended DEIM algorithm for subset selection and class identification*, Mach. Learn., 110(4) (2021), pp. 621–650.
6. I.T. Jolliffe: *Discarding variables in a principal component analysis. I: Artificial data*, Appl. Statist. 21(2) (1972), pp. 160–173.
7. M. W. Mahoney and P. Drineas: *CUR matrix decompositions for improved data analysis*, Proc. National Academy of Sciences 106(3) (2009), pp. 697–702.
8. D. Papailiopoulos, A. Kyrillidis and C. Boutsidis: *Provable deterministic leverage score sampling*, Proc. 20th ACM SIGKDD Conf. Knowl. Disc. Data Mining (2014), pp. 997–1006.
9. D.C. Sorensen and M. Embree: *A DEIM induced CUR factorization*, SIAM J. Sci. Comp., 38 (2016), pp. A1454–A1482.

Data-Driven Modeling and Control of Complex Dynamical Systems Arising in Renal Anemia Therapy

Sabrina Casper, Doris H. Fuertinger, Peter Kotanko, Luca Mechelli, Jan Rohleff, and Stefan Volkwein

Abstract This project is based on a mathematical model of erythropoiesis for anemia (Fuertinger, A model of erythropoiesis. PhD thesis, Karl-Franzens University Graz, 2012; Fuertinger et al., J Math Biol 66(6):1209–1240, 2013), which consists of five hyperbolic population equations describing the production of red blood cells under treatment with epoetin-alfa (EPO). Extended dynamic mode decomposition (EDMD) is utilized to approximate the non-linear dynamical systems by linear ones. This allows for efficient and reliable strategies based on a combination of EDMD and model predictive control (MPC), which produces results comparable with the one obtained in Rogg et al. (J Math Biol 79:2281–2313, 2019) for the original model.

1 Introduction

Almost all hemodialysis patients suffer from chronic anemia, due to the reduced functionality of the kidneys and the resulting low production of erythropoietin, a kidney-derived hormone that increases red blood cell output by the bone marrow. Therefore, physicians use erythropoietin stimulating agents, such as epoetin-alfa (EPO), to partially correct the anemia. The challenge in designing efficient therapies is due to the patients' differences in long-term response to EPO. In [1, 2], the authors

S. Casper · D. H. Fuertinger
Fresenius Medical Care Deutschland GmbH, Bad Homburg, Germany
e-mail: Sabrina.Rogg@fmc-ag.com; Doris.Fuertinger@fmc-ag.com

P. Kotanko
Renal Research Institute New York, New York, NY, USA
e-mail: Peter.Kotanko@rriny.com

L. Mechelli (✉) · J. Rohleff · S. Volkwein
Department of Mathematics and Statistics, University of Konstanz, Konstanz, Germany
e-mail: luca.mechelli@uni-konstanz.de; jan.rohleff@uni-konstanz.de; stefan.volkwein@uni-konstanz.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_21

introduce a mathematical model for predicting such a response. As in [7], our aim is to design a feedback control strategy, based on Model Predictive Control (MPC) [3], to optimize the injections of EPO doses in order to reach a target hemoglobin level. In contrast to [7], we do not imply the EPO model from [2] during the optimization, but we utilize it to generate a data-driven approximation of such a model through Extended Dynamic Mode Decomposition (EDMD) [8]. We then use such an EDMD-based surrogate model during the MPC. We refer to [5] where the authors introduced the idea to combine MPC and EDMD. We organize the work as follows: in Sect. 2 we introduce the model and how we combine EDMD and MPC. In Sect. 3 we show the reliability of the technique comparing our algorithm with the one in [7].

2 The EDMD-MPC Based Algorithm

Let us consider a finite horizon time interval $[0, T]$ with $T \gg 0$ and a finite number of injections $n_u \in \mathbb{N}$. We consider discrete injections of doses u_i at predefined injection times t_i for $i = 1, \dots, n_u$. Given a set of EPO doses $\mathbf{u} = (u_i)_{i=1}^{n_u} \in U_{\mathsf{ad}} = \{\mathbf{u} \in \mathbb{R}^{n_u} | 0 \le u_i \le u_{\max}, 1 \le i \le n_u\}$, following [7], one can compute the corresponding EPO concentration in the blood, which we indicate with $E(t; \mathbf{u})$. This nonlinear function is continuously differentiable in $[0, T]$ and twice continuously differentiable in U_{ad}. For further details see [2, 7]. The EPO model is composed of five coupled advection-reaction partial differential equations (PDEs) of the form

$$\begin{aligned} y_t(t,x) &= \kappa(x, E(t;\mathbf{u}))y(t,x) - v(E(t;\mathbf{u}))y_x(t,x) && \text{in } Q, \\ y(t,\underline{x}) &= g(t; E(t;\mathbf{u})) && \text{in } (0,T), \\ y(0,x) &= y_0(x) && \text{in } \Omega \end{aligned} \tag{1}$$

with the spatial domain $\Omega = (\underline{x}, \overline{x}) \subset \mathbb{R}$, the space-time cylinder $Q = (0, T) \times \Omega$ and the initial condition y_0. The solution $y(t, x)$ to (1) denotes the cell density of the respective cell population with maturity x at time t. The coupling among the five equations is hidden in the boundary value $g(t; E(t; \mathbf{u}))$. For the sake of brevity, we omit further explanations and we refer to [2, 7]. Note that in (1), the control $\mathbf{u}$ (EPO doses) enters in the equations through the nonlinear EPO concentration, on which the advection and reaction coefficients depend. Such a complicated relationship between the states (cell densities) y and the control $\mathbf{u}$ leads to a non-convex optimization problem. Our aim is then to introduce a linear surrogate model based on EDMD, such that the resulting control-to-state map is linear. Doing so, we have the advantage of obtaining a convex optimization problem, which admits a unique minimizer and can be solved faster. In contrast to the standard Dynamic Mode Decomposition (DMD) [6], EDMD utilizes snapshots of the dynamic, controls and observables of a dynamical system to extract a discrete surrogate model [8]. Let us mention that the EDMD method is a solely data-driven method, which is not related to the structure (or even the knowledge) of the model from where the data come from. For this project we used the original model to create a data set for the EDMD and to compare how well we can recover the data. In practice, a combination of clinical data (hemoglobin data) and model data (red blood cell (RBC) population in combination with an output model that computes the hemoglobin data from the total RBC population) would be used to inform and update the EDMD. Further we discretize in space with Legendre polynomials (as [7]) and in time with a constant time step Δt. In what follows n is the number of Legendre polynomials and m is the number of time steps. Let $\psi : \mathbb{R}^n \to \mathbb{R}^{N_\psi}$ be a vector of lifting functions (or observables) $\psi_i : \mathbb{R}^n \to \mathbb{R}$, i.e. $\psi(x) = \big(\psi_1(x), \ldots, \psi_{N_\psi}(x)\big) \in \mathbb{R}^{N_\psi}$. Let $Y_0 = [y_0|...|y_{m-1}] \in \mathbb{R}^{n\times m}$ and $Y_1 = [y_1|...|y_m] \in \mathbb{R}^{n\times m}$ and $U = [u_0|\ldots|u_{m-1}] \in \mathbb{R}^{1\times m}$ be given snapshots data matrices. Next, we need to define the matrices

$$Y_{0,\text{lift}} = \begin{pmatrix} \psi_1(y_0) & \ldots & \psi_1(y_{m-1}) \\ \vdots & & \vdots \\ \psi_{N_\psi}(y_0) & \ldots & \psi_{N_\psi}(y_{m-1}) \end{pmatrix}, \; Y_{1,\text{lift}} = \begin{pmatrix} \psi_1(y_1) & \ldots & \psi_1(y_m) \\ \vdots & & \vdots \\ \psi_{N_\psi}(y_1) & \ldots & \psi_{N_\psi}(y_m) \end{pmatrix} \in \mathbb{R}^{N_\psi\times m}$$

and identify the matrices $A \in \mathbb{R}^{N_\psi\times N_\psi}$, $B \in \mathbb{R}^{N_\psi\times n_u}$ and $C \in \mathbb{R}^{n\times N_\psi}$ such that

$$[A, B] = \underset{\tilde{A},\tilde{B}}{\operatorname{argmin}} \|Y_{1,\text{lift}} - \tilde{A}Y_{0,\text{lift}} - \tilde{B}U\|_F, \quad C = \underset{\tilde{C}}{\operatorname{argmin}} \|Y_0 - \tilde{C}Y_{0,\text{lift}}\|_F. \tag{2}$$

To solve (2) numerically, we have to perform two singular value decomposition; cf. [5]. We get then a discrete linear dynamical system in the observable space

$$\begin{aligned} z_{k+1} &= Az_k + Bu_k \text{ for } k \geq 0, \quad z_0 = \big(\psi_1(y_0), \ldots, \psi_{N_\psi}(y_0)\big), \\ \hat{y}_{k+1} &= Cz_{k+1} \text{ for } k \geq 0, \end{aligned} \tag{3}$$

where $\hat{y}$ is the EDMD approximation of the y solution to (1). Note that reconstructing and solving the EDMD system (3) is generally cheaper than solving the system of coupled hyperbolic PDEs (1). Thus, (3) is the surrogate model we will use during the MPC algorithm. More specifically, our goal is to apply EDMD to the fifth (and last) equation of the EPO model in [7], which has the structure of (1), for two different choices of control snapshots, i.e. the EDMD matrix U will contain:

(EDMD-C) The continuous EPO concentration $E(t; \mathbf{u})$ at each time step;
(EDMD-D) The discrete EPO doses $\mathbf{u}$ at the injection times and 0 for the rest.

The cost functional for the model in [7] discretized by Legendre polynomials is

$$\begin{aligned} J_N(\tilde{y}, \mathbf{u}) = \frac{1}{2}\sum_{j=1}^{n_u} \gamma_j u_j^2 + \frac{\sigma_\omega}{2}\int_0^T \Big(\omega_5^{1/2}(\tilde{y}_5)_0(t) - P^{\mathsf{d}}\Big)^2 \, \mathrm{d}t \\ + \frac{\sigma_f}{2}\Big(\omega_5^{1/2}(\tilde{y}_5)_0(T) - P^{\mathsf{d}}\Big)^2, \end{aligned}$$

Algorithm 2 EDMD-MPC algorithm

```
Data: Initial control $\mathbf{u}^{(0)} \in U_{\text{ad}}$, initial condition $y^{(0)}_{\text{MPC}}$, EDMD update tolerance $\tau_{\text{upd}}$, MPC prediction horizon $M$, EDMD update steps $M_{\text{EDMD}}$.
Initialize the EDMD model at $\mathbf{u}^{(0)}$ with snapshots from (1);
for $i = 0, 1, \dots$ do
    Solve (4) in $[t_i, t_i + M\Delta t]$ with projected BFGS and initial guess $y^{(i)}_{\text{MPC}}$ to get an optimal $\bar{\mathbf{u}}^{(i)}$;
    Store $u_{\text{MPC}}(t_i) = \bar{\mathbf{u}}^{(i)}(t_i)$ and compute the new initial guess $y^{(i+1)}_{\text{MPC}}$ from (1) using $u_{\text{MPC}}(t_i)$;
    if $\|y^{(i+1)}_{\text{MPC}} - y^{(i+1)}_{\text{EDMD}}\| > \tau_{\text{upd}}$ then
        Update the EDMD model using snapshots of (1) for $M_{\text{EDMD}}$ steps with the control $\bar{\mathbf{u}}^{(i)}$;
    end if
end for
```

where $(\cdot)_0$ means first-component in space and all the other parameters can be found in [7]. Moreover, $\tilde{y} = \mathcal{S}_N(\mathbf{u})$ is the solution of (24b) in [7]. We point out that J_N depends only on the solution of the fifth (and last) equation of the EPO model in [7]. Therefore, to build our EDMD model, we use the solution $\tilde{y}_5$ as snapshots of the dynamic. We then get matrices A, B and C for (3) according to the chosen snapshots. Applying a trapezoidal rule to J_N and considering the EDMD approximation $\hat{y}$, one obtains a cost functional $J_m(z, \mathbf{u}) = J_{N,\text{disc}}(Cz, \mathbf{u}) = J_{N,\text{disc}}(\hat{y}, \mathbf{u})$ and the optimal control problem is given as

$$\min J_m(z, \mathbf{u}) \quad \text{s.t.} \quad z = \{z_k\}_{k=0}^{m} \subset \mathbb{R}^{N_\psi} \text{ satisfy (3) for } \mathbf{u} \in U_{\text{ad}}. \tag{4}$$

Note that (4) is a convex linear-quadratic optimal control problem and thus admits a unique minimizer [4]. If the EDMD approximation error is small enough, the solution of (4) is in a neighborhood of a local minimizer of the optimal control problem in [7]. Since the horizon T is generally large, the corresponding time discretization $t_0 = 0, \dots, t_m = T$ contains many points. To avoid costly computations and compute a feedback control we introduce an MPC [3] framework. This technique consists in fixing a prediction horizon M and computing the solution of a first optimal control problem in $[0, M\Delta t]$, then storing the optimal control for the first time step, applying it to (1) and repeating the procedure for the horizon $[t_1, t_1 + M\Delta t]$ and so on. This approach has also the advantage of obtaining a feedback control which reacts to the solution of the EPO model. Since the EDMD is just a local approximation of the dynamics, we need to define a strategy in order to update the EDMD model during the MPC. We simply measure the difference between the EDMD solution and its EPO model counter part for the first time step. Note that this does not require additional computations with respect to the described procedure. We resume our algorithm in Algorithm 2.

3 Numerical Experiments

In this section, we compare the nonlinear MPC from [7] with the linear MPC based on the EDMD approach. For brevity, we report only the results on test conducted on Patient 2 and 3 of [7]. Let us mention, that we performed the numerical experiments also for the other patients in [7] and we got approximation errors similar to the one presented below. For both EDMD-MPC models, we choose an update tolerance $\tau_{\text{upd}} = 0.01$, $M_{\text{EDMD}} = 30$ steps, the MPC prediction horizon $M = 14$ days and

$$\psi_1(y) = \omega_5^{1/2}(y)_0, \quad \psi_{i,j}(y) = L^j\left(\frac{(y)_i}{\sum_{i=1}^{n}(y)_i}\right), \quad i = 0, \ldots, n,\ j = 0, \ldots N_L{-}1,$$

where $(y)_i$ is the i-th component of y and L^j is the j-th Legendre polynomial. It follows that $N_\psi = 1 + nN_L$. Note that the Legendre polynomials are not only used for the spatial discretization of (1), but also for defining the lifting functions $\psi_{i,j}$. These two sets of polynomials are anyway not related between each other. Furthermore, N_L can be changed arbitrarily leading to different approximation results for EDMD. In our case we fix $N_L = 2$ for EDMD-C and $N_L = 6$ for EDMD-D in order to get comparable accuracy for the two techniques. All the other parameters are chosen as in [7, Tables 4-8]. First of all, we consider a total time period of three weeks, i.e. $T = 21$ days. We have an injection at day 1, 3 and 5 of each week.

In Fig. 1-left panel, we plot the optimal hemoglobin level computed using the method from [7] (RRI MPC) and with the two EDMD-MPC approaches proposed in Sect. 2. As one can see, for the first part of the time horizon, the three methods compute exactly the same optimal solution. These corresponds to a doses $u = 0$, which is one of the constraint imposed on the doses. As soon as the control starts impacting the system, we can notice some differences arising between the RRI MPC and our proposed method, in particular for the EDMD-C approach. At each time step, such a difference remains approximately smaller than the 1%, as it can be seen from Fig. 1-right panel, where the relative error between RRI MPC and our method is reported. We observe similar results for Patient 3 as for Patient 2, see

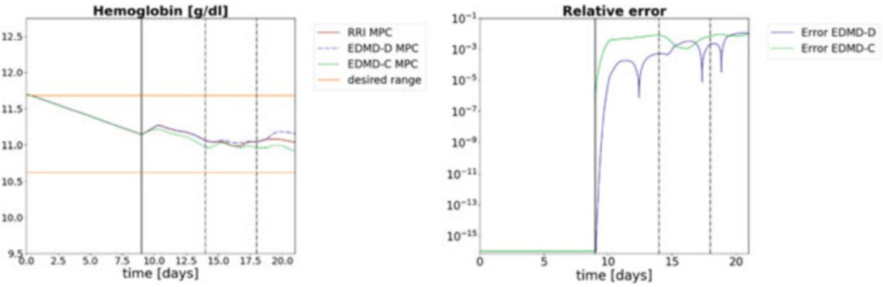


Fig. 1 Patient 2—21 days. **Left:** MPC solutions. **Right:** Relative error. (Black line: EDMD-D update, dashed black line: EDMD-C update)

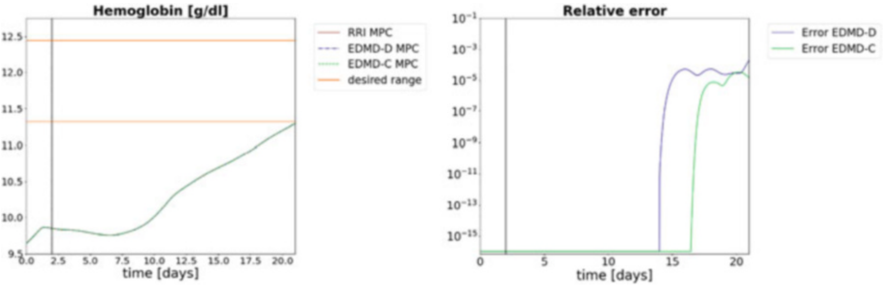


Fig. 2 Patient 3—21 days. **Left:** MPC solutions. **Right:** Relative error. (Black line: EDMD-D update, dashed black line: EDMD-C update)

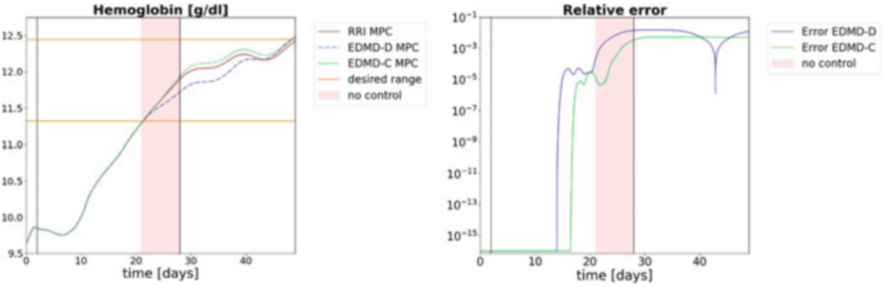


Fig. 3 Patient 3—49 days. **Left:** MPC solutions. **Right:** Relative error. (Black line: EDMD-D update, dashed black line: EDMD-C update, Red area: No injection possible)

Fig. 2 for further details. It appears that reconstructing the optimal hemoglobin level works extremely well (up to machine precision) for the first half of the time horizon and become worse in the second half (cf. Fig. 2). In this case, note that the EDMD updates are not triggered, even though the approximation worsens as time passes. This suggest that the chosen update strategy is too simple and its improvement will be object of future work. In the next test we consider a longer total time period of 7 weeks, i.e. $T = 49$ days, for Patient 3.

This test simulates that Patient 3 skips 1 week of injections for some reason. In Fig. 3 the week without injections is represented by the red area. For the MPC framework we do not require this additional information. We just assume that the patient will get three injections every week, until he skips the injection appointments. The test demonstrates the ability of the MPC algorithm to react with a quick feedback response. In Fig. 3 we see that our MPC methods based on EDMD are reacting as well (additionally updating the EDMD approximation) and their response is close to the non-linear MPC from [7]. For all the numerical tests, we report the space-time relative error in reconstructing the RRI MPC hemoglobin level and the required computational times for the proposed EDMD-MPC schemes in Table 1. Note that our methods are almost one order of magnitude faster and reconstruct the solution with a reliable approximation error. In conclusion, the proposed EDMD-MPC method replicates the results obtained in [7] with reasonable

Table 1 Computational time, speed-up w.r.t. [7] and relative error of the EDMD-MPC algorithm

Patient	Method	T	Computational time (including updates)	Speed-up	Relative error
2	EDMD-D	21 d	11.5 s	9.5	9.1×10^{-6}
2	EDMD-C	21 d	11.4 s	9.6	1.9×10^{-5}
3	EDMD-D	21 d	5.3 s	10.3	9.0×10^{-10}
3	EDMD-C	21 d	6.0 s	9.1	1.0×10^{-10}
3	EDMD-D	49 d	17.5 s	7.1	7.4×10^{-5}
3	EDMD-C	49 d	9.9 s	12.6	1.4×10^{-5}

error and a small factor of speed-up. This factor remains constant when the treatment horizon increases (cf. Table 1). The EDMD-MPC algorithm can be then a valid approach for hemodialysis treatments, although the estimates of the EDMD error and the resulting update strategy during the MPC iterations need to be improved. This will be the focus of a future work.

Acknowledgments L. Mechelli and J. Rohleff gratefully acknowledge support of the Independent Research Grant *Efficient model order reduction for model predictive control of non-linear input-output dynamical systems* of the Zukunftskolleg at the University of Konstanz. The authors thank Christian Himpe, Max Planck Institute Magdeburg, for the fruitful discussions and remarks.

References

1. Fuertinger, D.H.: A model of erythropoiesis. PhD thesis, Karl-Franzens University Graz (2012)
2. Fuertinger, D.H., Kappel, F., Thijssen, S., Levin, N.W., Kotanko, P.: A model of erythropoiesis in adults with sufficient iron availability. J. Math. Biol. **66**(6), 1209–1240 (2013)
3. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control:Theory and Algorithms. 2nd Edition. Springer, London (2016)
4. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Springer-Verlag, Berlin (2009)
5. Korda, M., Mezić I.: Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. Automatica **93**, 149–160 (2018)
6. Kutz, J.N., Brunton, S.L., Brunton, B.W., Proctor J.L.: Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems. SIAM, Philadelphia (2016)
7. Rogg, S., Fuertinger, D.H., Volkwein, S., Kappel, F., Kotanko, P.: Optimal EPO dosing in hemodialysis patients using a non-linear model predictive control approach. J. Math. Biol. **79**, 2281–2313 (2019)
8. Williams, M.O., Kevrekidis, I.G., Rowley, C.W.: A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. J. Nonlinear Sci. **25**, 1307–1346 (2015)

Regional Estimates of Reproduction Numbers with Application to COVID-19

Stefan Heyder, Jan Pablo Burgard, Tyll Krueger, and Thomas Hotz

Abstract In the last year many public health decisions were based on real-time monitoring the spread of the ongoing COVID-19 pandemic. For this one often considers the reproduction number which measures the amount of secondary cases produced by a single infectious individual. While estimates of this quantity are readily available on the national level, subnational estimates, e.g. on the county level, pose more difficulties since only few incidences occur there. However, as countermeasures to the pandemic are usually enforced on the subnational level, such estimates are of great interest to assess the efficacy of the measures taken, and to guide future policy. We present a novel extension of the well established estimator (Fraser, PloS One 2:8, 2007) of the country level reproduction number to the county level by applying techniques from small-area estimation. This new estimator yields sensible estimates of reproduction numbers both on the country and county level. It can handle low and highly variable case counts on the county level, and may be used to distinguish local outbreaks from more widespread ones. We demonstrate the capabilities of our novel estimator by a simulation study and by applying the estimator to German case data.

S. Heyder (✉) · T. Hotz
Technische Universität Ilmenau, Ilmenau, Germany
e-mail: stefan.heyder@tu-ilmenau.de; thomas.hotz@tu-ilmenau.de

J. P. Burgard
Universität Trier, Trier, Germany
e-mail: burgardj@uni-trier.de

T. Krueger
Wroclaw University of Science and Technology, Wrocław, Poland
e-mail: tyll.krueger@pwr.wroc.pl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_22

1 Introduction

The ongoing COVID-19 pandemic is affecting countries worldwide with over 4.4 million deaths as of 30 August 2021 [8]. To restrict the spread of SARS-CoV-2, the virus causing COVID-19, many countries have implemented non-pharmaceutical countermeasures such as bans of mass gatherings, mandatory wearing of masks and reduction of contacts in the private and work life. In addition vaccines which reduce both the severity of COVID-19 and the infectiousness of vaccinated individuals have become available, and most European countries have vaccinated large portions of their population [1].

To quantify the spread of an epidemic, one considers the time-varying reproduction number $R(t)$, the mean amount of secondary cases a primary case infected on day t is expected to infect during his course of infection, provided conditions stay the same. Knowing $R(t)$ allows one to infer whether the number of cases will rise or fall in the future; the threshold for growth being $R(t) = 1$. On the country-level a standard model for the spread of an epidemic is the following stochastic renewal equation for $I(t)$, the amount of newly infected cases on day t, which are assumed to be (conditionally) Poisson distributed:

$$I(t) \mid I(t-1), \dots \sim \text{Pois}\left(R(t) \sum_{\tau=1}^{\infty} I(t-\tau) w(\tau) \right). \tag{1}$$

Here, $w(\cdot)$ specifies the distribution of the generation time, i.e., given that a primary case infects a secondary case, $w(\tau)$ is the probability that this infection occurs on day τ after the primary case was infected himself. A well studied estimator of $R(t)$ in this model is

$$\hat{R}(t) = I(t) / \sum_{\tau=1}^{\infty} I(t-\tau) w(\tau) \ , \tag{2}$$

see e.g. [2, 4]. For this estimator to be reliable the denominator has to be large enough, as its variance (conditional on past cases) is $R(t)/\sum_{\tau=1}^{\infty} I(t-\tau)w(\tau)$, see [5].

A deficit of estimating the reproduction number on the country level is that these estimates are affected by local outbreaks which, in the absence of high case numbers, dominate even country-level estimations. In the reproduction number estimation this causes undesirable artifacts: the nationwide spread of the epidemic is first overestimated due to the local outbreak while later the country-wide reproduction number will be underestimated since the denominator of $\hat{R}(t)$ is too large due to the previous outbreak, for example Fig. 1 shows the effect of a huge influx of cases in June 2020 in Germany due to several smaller outbreaks, the biggest with 1413 cases occuring in a meat processing plant in Gütersloh county [3].

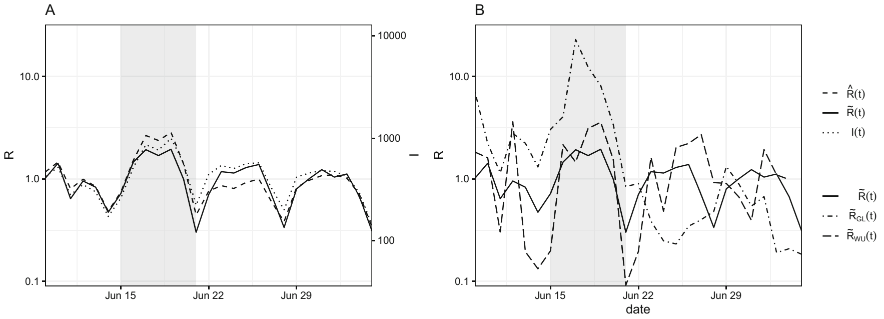


Fig. 1 (**a**) shows reproduction number estimates and reported cases (*dotted line*), both on a logarithmic scale, in Germany. On 17 June 2020 the first cases of a local outbreak were reported, causing a spike in the estimated reproduction numbers. Another consequence of this outbreak are lower estimates of the reproduction numbers (*dashed line*) in the following weeks. Both phenomena are less pronounced for the estimate based on county level data (*solid line*). (**b**) additionally shows county-level reproduction number estimates of Gütersloh county, $\tilde{R}_{\mathrm{GL}}(t)$ (*dot-dashed line*), and Wuppertal county, $\tilde{R}_{\mathrm{WU}}(t)$ (*double-dashed line*)

Small area estimation (SAE) is a branch of mathematical statistics providing tools suited for precisely this situation: data per region are scarce and may even be missing but there are many regions. To make a virtue out of necessity, SAE models regional parameters as random variables, an approach we apply to county-level reproduction numbers. Specifying the joint distribution of county-level reproduction numbers enables us to estimate a single set of parameters from which we can compute an estimated distribution of the reproduction number in each county. This procedure can be viewed as empirical Bayes estimation. We show that reproduction numbers obtained this way can be used to identify local outbreaks, handle low case numbers while agreeing with the country level estimates of the reproduction number [5] in the absence of local outbreaks.

2 Estimator

A standard way of modeling the infection process is the renewal equation (1), cf. [2] for a detailed derivation. We present a straight forward generalization of this model to the regional level by using techniques from small-area estimation. In small-area estimation it is common to model parameters on the regional level to vary randomly; in this spirit we model $R_c(t)$, the regional reproduction number on day t in region c, by a random variable.

To account for cases that are imported and exported between regions, we assume that a fraction p_t of secondary cases are attributed to a region different than the corresponding primary case. Let $\Phi_c(t) = \sum_{\tau=1}^{\infty} I_c(t-\tau)w(\tau)$ be the expected number of active cases on day t in county c given the past where $I_c(t)$ denotes the incidences in that region on that day. We then use the following renewal equation to describes the spread of the epidemic, relating the conditional distribution of $I_c(t)$ to the expected number of active cases and the regional reproduction number $R_c(t)$:

$$I_c(t)|R_c(t), I_c(t-1), I_{c'}(t-1)\cdots \sim \text{Pois}\big(R_c(t)\big((1-p_t)\Phi_c(t) + \tfrac{p_t}{K-1}\sum_{c'\neq c}\Phi_{c'}(t)\big)\big) \tag{3}$$

Here K denotes the total number of regions considered. Note that we condition not only on past incidences $I_c(t-\tau)$ in all counties but also on the random reproduction number $R_c(t)$.

The interpretation of (3) is straight-forward: on day t there are $I_c(t-\tau)$ individuals τ days into their infection, thus $R_c(t)w(\tau)I_c(t-\tau)$ is the expected amount of secondary infections caused by these individuals on day t. To account for the transfer of cases between counties, a fraction of p_t cases are counted towards the active cases in other regions and the wrongfully attributed cases are distributed equally among all other regions. Summing over τ yields the new infections $I_c(t)$ which we assume to be Poisson distributed.

To infer $R_c(t)$ from (3) further assumptions about both the distribution of $R_c(t)$ and the joint distribution of the pairs $(I_c(t), R_c(t))$ for all regions c are necessary. To this end we assume that the regional reproduction numbers on day t posses a common, known distribution and that the set of tuples $(I_c(t), R_c(t))$ is conditionally independent (given past incidences). More concretely, we assume the common distribution of the regional reproduction numbers $R_c(t)$ to be a gamma distribution $\text{Gamma}(a_t, s_t)$ with *shape* a_t and *scale* s_t and density $\frac{1}{s_t^{a_t}\Gamma(a_t)}x^{a_t-1}\exp\left(\frac{-x}{s_t}\right)$.

It is easy to see that the marginal distribution of $I_c(t)$ (given the past incidences) in that region—without conditioning on the reproduction number $R_c(t)$—, is then a mixture of a gamma and a Poisson distribution, i.e. a negative binomial distribution whose parameters only depend on the parameters a_t, s_t, p_t, past incidences $I_c(t-\tau)$ and the generation time distribution w.

As the conditional distribution of $I_c(t)$ only depends on the unknown parameters, and, conditionally, the incidences of different regions are independent, we can apply maximum-likelihood estimation to obtain estimates $\hat{a}_t$, $\hat{s}_t$ and $\hat{p}_t$ of the unknown parameters.

Also, the gamma distribution is conjugate prior to the Poisson distribution whence the conditional distribution of $R_c(t)$ given past incidences $I_c(t), I_c(t-1), \ldots$ is again a gamma distribution whose shape and scale only depend on the unknown parameters and past incidences. Thus one can use plug-in to estimate parameters of the posterior distribution such as $\mathbf{E}(R_c(t)|I_c(t), \ldots)$ and to derive prediction intervals. Furthermore we naturally obtain a new estimator of the country-wide reproduction number, the estimated mean $\tilde{R}(t) = \hat{a}_t\hat{s}_t$.

This approach could also be interpreted in the setting of empirical Bayes methods if one thinks of $\text{Gamma}(a_t, s_t)$ as the prior distribution of $R_c(t)$ and $I_c(t)$ as the observations, with the prior parameters being estimated with tools from frequentist statistics.

3 Parameters, Data Sources and Implementation Details

The estimators consider assume the probability mass function w of the generation time to be known. As a precise model for the generation time is difficult to obtain we opt for a simple model: we assume the shape of w to be trapezoidal with a mean of 5.6 days in accordance with the mean serial interval of 5.4 days found in [9], see [5] for details. In the same spirit we assume that the generation time distribution does not change over time.

To estimate the county-level reproduction numbers in Germany we use data provided by the Robert-Koch Institut [7], as of 30 August 2021. This dataset contains daily information on reported cases and deaths in Germany in addition to the county (Landkreis) where the case was reported to local health authorities. There is a strong weekday effect present in both the case and death counts. This effect is most likely due to testing, evaluating tests and reporting occuring more frequently on workdays compared to weekends. We do not account for this effect to direct the readers attention to the existence of such artifacts in the data and to avoid overconfidence in the resulting estimates—these should be interpreted qualitatively not quantitatively.

Note that there is a delay between infection and reporting of cases so that estimates of reproduction numbers $\hat{R}(t)$, $\tilde{R}(t)$ ought to be backdated by about 7 days, see [5] for details.

All computations, including simulations to validate the estimator, are conducted in `R` version 4.1.1 [6]. The calculation of maximum-likelihood estimates $\hat{a}_t$, $\hat{s}_t$, $\hat{p}_t$ cannot be performed analytically, and is achieved using numerical optimization by the built-in function `optim`.

4 Validation by Simulation

To check how a mismatch between our model and reality might affect our estimator, we simulate a point process on the flat torus $\mathbf{T} = \mathbf{R}^2/(k\mathbf{Z})^2$, $k \in \mathbf{N}$ where each of the k^2 unit squares corresponds to a county. We chose $k = 20$ to obtain $k^2 = 400$ counties, approximating the 401 counties in Germany. Time is chosen to be discrete and measured in days. To simplify computation we simulate on $\mathbf{R}^2$ and quotient out $(k\mathbf{Z})^2$ after the simulation has finished.

We initialize the simulation with 400 infected individuals that are placed uniformly on $\mathbf{T}$, their infection age chosen again uniformly from the discrete support of the trapezoidal generation time distribution w (see Sect. 3). At each time t every infected individual with infection age τ in county c infects a random, $\text{Pois}(R_c(t)w(\tau))$-distributed, number of new cases.

The position of the new cases is also random, and sampled from a bivariate normal distribution centered at the position of the primary case with covariance matrix $\sigma^2\mathbf{I}_2$. We chose σ^2 such that approximately 20% of secondary cases occur in counties different from their primary case, resulting in $\sigma^2 \approx (0.14)^2$.

These simulations introduce a mismatch between model (3) and the generated incidence data. Firstly, exported cases are no longer distributed evenly over all counties, but rather depending on proximity. Secondly, we can choose the reproduction numbers to deviate from the assumed Gamma distribution. To incorporate the introduction and partial lifting of non-pharmaceutical interventions we set $R(t)$ to be 2.5 for 20 days, 0.7 for 40 days and 1.2 for another 40 days, simulating an outbreak over a total of 100 days.

The daily reproduction number estimates based on the case data of this simulation as well as asymptotic 95% confidence sets, based on the Fisher information, are shown in Fig. 2a. Despite the model mismatch the coverage of the confidence intervals is close to 95% and also stays this way if we simulate this scenario multiple times (figures not shown). Additionally the sharp changes in the reproduction number on days 21 and 61 are captured by our estimator as well.

We also show an estimate of $\mathbf{E}(R_c(t)|I_c(t),\dots)$, the county level reproduction numbers, for every county in Fig. 2b. In this model the county level reproduction numbers have zero variance. This results in some estimates of the variance $a_t s_t^2$ to be very small, making all county level estimates similar at some time points. Increasing the regional variation by sampling reproduction numbers from a Gamma distribution did not produce such effects (figures not shown).

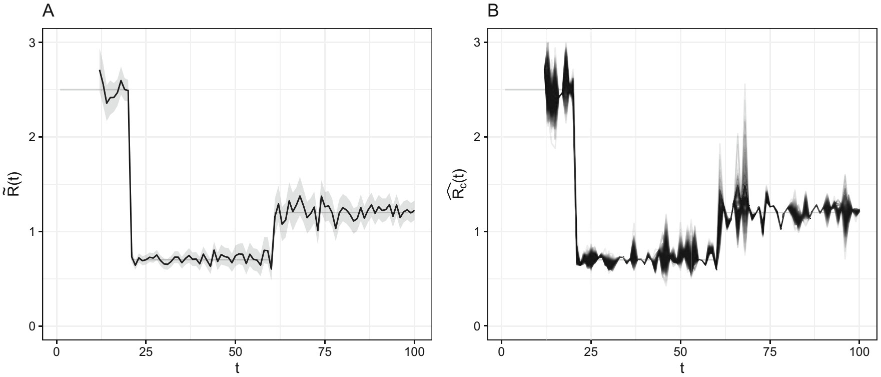


Fig. 2 Results for one simulated outbreak. (**a**) shows the estimates of the posterior mean $\hat{a}_t\hat{s}_t$ in *black* with corresponding confidence intervals indicated by *grey ribbons*, the true $R(t)$ is shown as a *transparent grey line*. (**b**) shows the estimates of reproduction numbers on the county level

5 Application to the COVID-19 Pandemic in Germany

In Fig. 1 we depict our new estimator $\tilde{R}(t)$ with $\hat{R}(t)$ for Germany, with a special focus on the aforementioned outbreak in June 2020. The weekly pattern in the estimates is due to the similar pattern in the incidence data; we decided against smoothing the estimates to highlight these complications with the data quality. Note that in the week corresponding to the outbreak, $\tilde{R}(t)$ is lower than the previous estimate. Additionally, the downwards trend of $\hat{R}(t)$ in the following weeks with estimates below 1 is no longer present, as the outbreak was a local one in few counties. Except for the deviations mentioned above, $\tilde{R}(t)$ resembles $\hat{R}(t)$ remarkably well. Around October 2020 a second wave of infections started to occur in Germany with rapidly rising case numbers across the country. Figure 1 shows that under these circumstances, i.e., high incidences in all regions, the country level estimates based on the small area estimation approach do not differ much from the estimates based on the country level.

6 Discussion

Of course our estimator rests on assumptions which ought to be discussed. Modeling $R_c(t)$ as random is a standard approach in small area estimation when dealing with few or even missing observations on a sub-national level; it is required to reduce the dimensionality of the parameter space. For this, we critically assumed that on a fixed day t the regional reproduction numbers $R_c(t)$ in different counties are independent and identically distributed according to a gamma distribution. This is questionable as transmission dynamics vary with local social and economic factors. For example one might expect that reproduction numbers are higher in urban counties than in rural counties with less population density. Furthermore neighboring counties might exhibit spatial correlation. Such socio-economic factors might be incorporated as for generalized linear mixed effects models although it is not obvious which factors to include and how to model their influence on the parameters a_t, s_t and p_t.

Assuming a gamma distribution for the regional reproduction numbers $R_c(t)$ is mathematically convenient as it is the conjugate prior distribution to the Poisson distribution, so using plug-in to obtain estimates for the posterior parameters is easy. In addition the log-likelihood of the posterior predictive distribution can be calculated analytically which makes estimation fast. The price we pay for this distributional assumption is that the gamma distribution is a relatively light-tailed distribution prohibiting it from fully incorporating superspreading events such as the investigated outbreak. For this outbreak the country level estimates provided by $\tilde{R}(t)$ are still elevated when compared to the previous and next week (see Fig. 1), which might be an artifact of our choice of distribution as well as the small-area approach which biases estimates towards the country-wide mean. Changing the marginal distribution of $R_c(t)$ would lead to a computationally more involved estimation procedure requiring numerical integration. The results in Sect. 4, however, show that our estimators are rather robust against slight misspecification in the prior distribution.

In addition to the mathematical assumptions discussed above we also made some more subtle epidemiological assumptions. To account for infections across regions we introduced the parameter p_t, the proportion of cases that were attributed to a different region than the one where infection occurred. The addition of p_t is essential to the model when considering periods where incidence is low, e.g. during the summer in Germany. Without modeling cross-county infections, counties which have reached incidence 0 for a prolonged period of time would never record new cases, and observing new cases in such a county would lead to a breakdown of the estimator as the observed data would have likelihood 0. We assumed that such transferred infections spread evenly among the other counties and that the this spread is the same for all counties, though the results of Sect. 4 suggest robustness against such a model mismatch. This could be improved by spatial models for the transfer of cases, e.g. based on mobility data.

We also assume the generation time distribution w to be constant over time and to be known. The sensitivity of our new estimator to misspecification in the generation time could easily be studied by adapting the simulations from Sect. 4 to include such a mismatch between simulation and estimation. As this sensitivity is not the main concern of this paper, we omit such an analysis but refer the reader to [5].

We caution the reader to interpret the estimations and predictions proposed in this paper quantitatively due to the restrictions mentioned above as well as the quality of the available data. Nevertheless we believe that the presented estimation procedure can be used to yield qualitative insight about the behavior of sub-national spread of an epidemic when case counts are low. In such scenarios our estimator $\tilde{R}(t)$ is a better representation of the country-level spread of the epidemic because it is less affected by local outbreaks.

References

1. European Centre for Disease Prevention and Control. COVID-19 Vaccine Tracker. https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html, 2021. Last accessed 30 August 2021.
2. Fraser, C. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one 2*, 8 (2007).
3. Günther, T., Czech-Sioli, M., Indenbirken, D., Robitaille, A., Tenhaken, P., Exner, M., Ottinger, M., Fischer, N., Grundhoff, A., and Brinkmann, M. M. SARS-CoV-2 outbreak investigation in a German meat processing plant. *EMBO Molecular Medicine 12*, 12 (Dec. 2020), e13296. Publisher: John Wiley & Sons, Ltd.
4. Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., and De Salazar, P. M. Practical considerations for measuring the effective reproductive number, R_t. *PLoS computational biology 16*, 12 (2020), e1008409.
5. Hotz, T., Glock, M., Heyder, S., Semper, S., Böhle, A., and Krämer, A. Monitoring the spread of COVID-19 by estimating reproduction numbers over time. *arXiv:2004.08557 [q-bio, stat]* (Apr. 2020).

6. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
7. Robert Koch-Institut. RKI COVID19. https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0, 2021. Last accessed 30 August 2021.
8. World Health Organization. COVID-19 Weekly Epidemiological Update Edition 54. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210824_weekly_epi_update_54.pdf, 2021. Last accessed 30 August 2021.
9. Zhang, P., Wang, T., and Xie, S. X. Meta-analysis of several epidemic characteristics of COVID-19. *Journal of data science: JDS 18*, 3 (July 2020).

Complexity Reduction for Parametric High Dimensional Models in the Analysis of Financial Risk

Andreas Binder, Onkar Jadhav, and Volker Mehrmann

Abstract This paper presents a parametric model order reduction (pMOR) approach for financial risk analysis based on the proper orthogonal decomposition method. pMOR requires solving the high dimensional model for some training parameters to obtain the reduced basis. We propose an adaptive greedy sampling approach based on surrogate modeling for the selection of training parameters. The developed algorithms are tested on an industrial example of a puttable steepener. The results show that the reduced model works excellent and has potential applications in historical or Monte-Carlo value at risk calculations.

1 Introduction

The risk analysis of financial instruments often requires the valuation of such instruments under a wide range of future market scenarios. In this paper, we use convection-diffusion-reaction partial differential equations (PDEs) of the interest rate models as valuation functions [1]. These models are usually calibrated based on market data like *yield curves* that generate a high dimensional parameter space [3]. In short, to perform the risk analysis, the financial model is needed to be solved for such a high dimensional parameter space, which requires efficient algorithms.

Thus, this paper establishes a parametric model order reduction (MOR) approach based on the proper orthogonal decomposition. This MOR approach requires simulating the high dimensional parametric model for a small number of pre-selected training parameter values. To select the training parameters, we employ an adaptive greedy sampling algorithm that constructs a surrogate model for the

A. Binder (✉)
MathConsult GmbH, Linz, Austria
e-mail: andreas.binder@mathconsult.co.at

O. Jadhav · V. Mehrmann
Institut für Mathematik, Berlin, Germany
e-mail: jadhav@math.tu-berlin.de; mehrmann@math.tu-berlin.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_23

error estimator and locates training parameters efficiently [2, 8]. To summarize, instead of performing thousands of costly simulations, we perform few expensive computations and simulate the remaining scenarios with the help of the reduced model.

We apply the MOR technique to packaged retail and insurance-based investment products (PRIIPs). In order to make PRIIPs from different manufacturers more comparable concerning their risks and returns, the European regulation (EU) 1286/2014 requires manufacturers of PRIIPs to supply key information documents (KIDs) to possible retail investors that are easy to read and to understand [4]. As per regulations, for long-term structured interest rate instruments under PRIIPs (maturity greater than 5 years), 10,000 computations must be carried out. Assuming a single valuation takes only one second, the whole computation will take almost a day. Therefore, it is interesting to test the developed model order reduction framework against such a challenging problem. We solve a numerical example of the puttable steepener with caps and floors using the two-factor Hull-White model.

2 Model Order Reduction

Consider a financial instrument $V(t, r(t), u(t))$ contingent on the stochastic interest rates movement $r(t)$, $u(t)$, the two-factor Hull-White PDE is then given as [6]

$$\begin{aligned}\frac{\partial V}{\partial t} + (\theta(t) + u - \alpha r)\frac{\partial V}{\partial r} - bu\frac{\partial V}{\partial u} + \frac{1}{2}\sigma_1^2\frac{\partial^2 V}{\partial r^2}\\ + \frac{1}{2}\sigma_2^2\frac{\partial^2 V}{\partial u^2} + \gamma\sigma_1\sigma_2\frac{\partial V}{\partial r\partial u} - rV = 0,\end{aligned} \tag{1}$$

where the deterministic function $\theta(t)$ is chosen to fit the simulated yield curves, parameters $\alpha, b, \sigma_1, \sigma_2 > 0$ are positive constants, t is time, and $-1 \leq \gamma \leq 1$. According to the PRIIPs regulations, we have to perform at least $s = 10{,}000$ yield curve simulations. We construct a simulated yield curve matrix $Y = [y_{s1}, \ldots, y_{sm}] \in \mathbb{R}^{s\times m}$, which is used to calibrate the parameter $\theta(t)$, where m is the number of yield curve *tenor/time points*. The calibration generates s different piecewise constant parameters $\theta_\ell(t)$, which change their values $\theta_{\ell,i}$ only at the m tenor points [2].

We have applied a finite element method to solve the PDE numerically, cf. [1]. This discretization is a parametric high dimensional model of the form

$$A(\rho_\ell(t))V^{n-1} = B(\rho_\ell(t))V^n, \tag{2}$$

with given terminal vector V^T, and matrices $A(\rho_\ell) \in \mathbb{R}^{M\times M}$, and $B(\rho_\ell) \in \mathbb{R}^{M\times M}$. $\rho = \{\alpha, b, \sigma_1, \sigma_2, \gamma, \theta(t)\}$ is the group of model parameters. We call this the *full model* (FM) for the model reduction procedure. We solve (2) by propagating backward in time. Altogether we have a parameter space $\mathcal{P}$ of size 10,000 × m to which we now apply model reduction.

To perform the parametric model reduction for system (2), we employ Galerkin projection onto a low dimensional subspace via

$$\bar{V}^n = QV_d^n, \tag{3}$$

where the columns of $Q \in \mathbb{R}^{M \times d}$ represent the reduced basis with $d \ll M$, V_d^n is a vector of reduced coordinates, and $\bar{V}^n \in \mathbb{R}^M$ is the solution in the nth time step obtained using the reduced model. For the Galerkin projection we require that the residual of the reduced state

$$p^n(V_d^n, \rho_\ell) = A(\rho_\ell)QV_d^{n-1} - B(\rho_\ell)QV_d^n \tag{4}$$

is orthogonal to the reduced basis matrix Q, which gives

$$\begin{aligned} Q^T A(\rho_\ell)QV_d^{n-1} &= Q^T B(\rho_\ell)QV_d^n, \\ A_d(\rho_\ell)V_d^{n-1} &= B_d(\rho_\ell)V_d^n, \end{aligned} \tag{5}$$

where $A_d(\rho_\ell) \in \mathbb{R}^{d \times d}$ and $B_d(\rho_\ell) \in \mathbb{R}^{d \times d}$ are the parameter dependent reduced matrices. We obtain the reduced basis Q in (3) using the method of snapshots that solves the full model for some training parameter groups $\{\rho_1, \ldots, \rho_k\}$ to generate a snapshot matrix $\hat{V} = [V(\rho_1), V(\rho_2), \ldots, V(\rho_k)]$. The POD method then solves, see [5], for an orthogonal matrix $Q \in \mathbb{R}^{M \times d}$ via a *truncated SVD*

$$\hat{V} = \Phi\Sigma\Psi^T = \sum_{i=1}^{k} \Sigma_i \phi_i \psi_i^T, \tag{6}$$

where ϕ_i and ψ_i are the left and right singular vectors of the matrix $\hat{V}$ respectively, and Σ_i are the singular values arranged in descending order. We choose only those d out of k left singular vectors to construct $Q = [\phi_1 \cdots \phi_d]$ which minimize the projection error, [7],

$$\epsilon_{\text{POD}} = \frac{1}{k}\sum_{j=1}^{k} \|V_j(\rho_j) - \sum_{i=1}^{d}(V_j(\rho_j)\phi_i)\phi_i\|^2 = \sum_{\ell=d+1}^{k} \Sigma_\ell^2. \tag{7}$$

It is evident that the quality of the reduced model strongly depends on the selection of parameter groups $\rho_1, \ldots, \rho_k$ that are used to compute the snapshots. Thus, we implemented an adaptive greedy sampling approach.

The basic idea of the greedy approach is to select the parameter groups at which the relative error between the reduced model and the full model is maximal. Thus, adding the full model solution for the worst parameter group to the snapshot matrix will ultimately improve the quality of the reduced basis for the next iteration. Since the relative error demands a costly full model solution, we replace it with an error estimator like a residual error $\varepsilon = \|p(.,\rho)\|$. In short, at each greedy iteration, the algorithm locates a parameter group that maximizes the residual error. However, it is not computationally feasible to compute an error estimator for the entire parameter space only to train the greedy algorithm. This problem forces us to randomly select a pre-defined parameter set $\hat{\mathcal{P}}$ as a subset of the parameter space $\mathcal{P}$ to train the greedy sampling algorithm. However, a random selection may neglect the crucial parameters within the parameter space. Thus, to surmount this problem, we implemented an adaptive greedy sampling approach, which selects these optimal parameter groups adaptively at each iteration of the greedy procedure, using an optimized search based on surrogate modeling [2].

The first stage of the adaptive greedy sampling Algorithm 1 computes the error estimator over a randomly selected parameter set $\hat{\mathcal{P}}_0$ of cardinality c_0. The algorithm uses these error estimator values $\{\varepsilon_i\}_{i=1}^{c_0}$ to build a surrogate model $\bar{\varepsilon}_0$. We solve this surrogate model for the entire parameter space and locate the c_k parameter groups corresponding to the c_k maximal values of the surrogate model. We construct a new

Algorithm 1 Adaptive greedy sampling algorithm

```
Input: Maximal number of iterations $I_{max}$, maximal number of parameter groups $c$, number of
    adaptive candidates $c_k$, parameter space $\mathcal{P}$, tolerance $\varepsilon_{tol}$.
Output: $Q$
Choose a parameter group $\rho_1$ from $\mathcal{P}$, simulate the full model for $\rho_1$ and store results in $V_1$
Compute a truncated SVD of the matrix $V_1$ and construct $Q_1$
for $i = 2, \dots, I_{max}$ do
    Randomly select a set of parameter groups $\hat{\mathcal{P}}_0 = \{\rho_1, \rho_2, \dots, \rho_{c_0}\} \subset \mathcal{P}$
    Compute error estimator values $\varepsilon(\rho_j)_{j=1}^{c_0}$ for each $\rho_j$ of $\hat{\mathcal{P}}_0$
    Let $\hat{\varepsilon}_0 = \{\varepsilon(\rho_1), \dots, \varepsilon(\rho_{c_0})\}$ be error estimators for $\hat{\mathcal{P}}_0$
    Let $k = 1$ and $e_{sg} = \hat{\varepsilon}_0$
    while $n(\hat{P}) < c$ do
        Construct a surrogate model $\bar{\varepsilon}(\rho)$ using the values $e_{sg}$ and solve it for $\mathcal{P}$
        Determine first $c_k$ maximal values of $\bar{\varepsilon}(\rho)$ and the parameter set $\hat{\mathcal{P}}_k = \{\rho_1, \dots, \rho_{c_k}\}$
        Compute error estimator values $\hat{\varepsilon}_k = \varepsilon(\rho_x)_{x=1}^{c_k}$ for each $\rho_x$ of $\hat{\mathcal{P}}_k$
        Update $e_{sg} = \{\hat{\varepsilon}_0 \cup \cdots \cup \hat{\varepsilon}_k\}$
        Construct a new parameter set $\hat{\mathcal{P}} = \hat{\mathcal{P}}_0 \cup \hat{\mathcal{P}}_k$
        $k = k + 1$
    end while
    Find $\rho_I = \underset{\rho \in \hat{\mathcal{P}}}{\text{argmax}}\ \varepsilon(\rho)$
    if $\varepsilon(\rho_I) \leq \varepsilon_{tol}$ then
        $Q = Q_{i-1}$
        break
    end if
    Solve the full model for $\rho_I$ and store results in $V_i$
    Construct a snapshot matrix $\hat{V}$ by concatenating the solutions $V_\ell$ for $\ell = 1, \dots, i$
    Compute a truncated SVD of the matrix $\hat{V}$ and construct $Q_i$
end for
```

parameter set $\hat{\mathcal{P}}_k = \{\rho_1, \dots, \rho_{c_k}\}$ composed of these c_k parameter groups. Once again, for each parameter group within the parameter set $\hat{\mathcal{P}}_k$, the algorithm simulates the reduced models, computes the error estimator values $\{\varepsilon_i\}_{i=1}^{c_k}$, and generates a new surrogate model $\bar{\varepsilon}_k$. This process repeats itself for $k = 1, \dots, K$ iterations until the total number of parameter groups reaches c. Finally, the optimal parameter group ρ_I is the one that maximizes the error estimator within the parameter set

$$\hat{\mathcal{P}} = \hat{\mathcal{P}}_0 \cup \hat{\mathcal{P}}_1 \cup \hat{\mathcal{P}}_2 \cup \cdots \cup \hat{\mathcal{P}}_K.$$

The algorithm then truncates after I_{max} iterations or until the maximum value of the error estimator drops below the tolerance. We have defined a surrogate model for a parameter group $\rho = [\rho_{\ell.1}, \dots, \rho_{\ell,m}]$ based on the principal component regression as follows, see [2],

$$\bar{\varepsilon}(\rho_\ell) = \eta_1 \rho_{\ell,1} + \cdots + \eta_m \rho_{\ell,m},$$

where $\eta_1, \dots, \eta_m$ are the regression coefficients.

3 Numerical Example

We consider a puttable steepener instrument whose coupons depend on the difference between two *constant maturity swap* (CMS) rates as a test example [3]. The coupons between years 1 to 3 are fixed at 4%, while the coupons from year 4 till maturity depend on (CMS10–CMS2). The coupon frequency is annually with a cap rate 3.0% and a floor rate 0.0% and the maturity is 10 years. It is a puttable type of instrument, i.e., a put is an option that gives the right to a buyer to sell the underlying asset at an agreed price at some time point in the future. In this paper, we have considered Put Price = 1. Figure 1 shows the monotonically decreasing projection error ϵ_{POD} associated with the proper orthogonal decomposition. The monotonically decreasing graph of the projection error shows that we have succeeded in determining a very good reduced basis. The relative error plot in Fig. 1 for a randomly selected parameter group ρ_{342} indicates that the error decreases with increasing reduced dimension d. We can see that the reduced model of the dimension $d = 10$ is satisfactory as the relative error is of order 10^{-4}. We solve this reduced model for 10,000 parameter groups and procure 10,000 corresponding values for the instrument, which are then sorted into three performance scenarios (favorable, moderate, unfavorable at 90th, 50th, 10th percentile), as shown in Fig. 2. We also noticed that the reduced model obtained using the adaptive greedy sampling approach is 8–10 times faster than the full model.

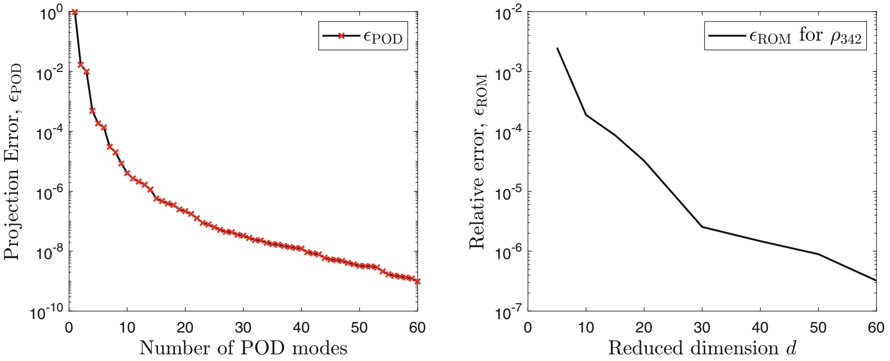


Fig. 1 Projection error associated with the proper orthogonal decomposition (left), and a plot of the relative error between the full model and the reduced model (right)

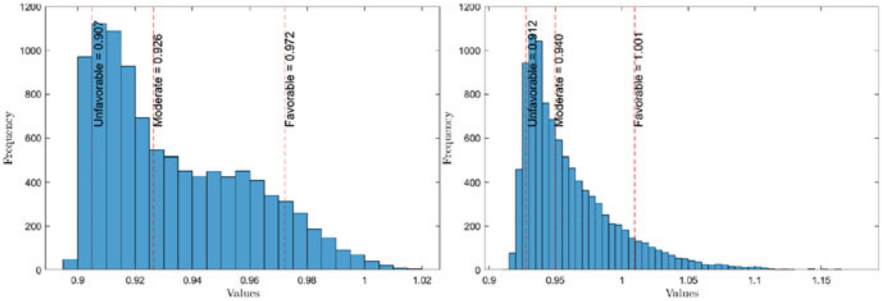


Fig. 2 Distribution of 10,000 results after five years (left) and ten years (right)

4 Conclusion

The results indicate the computational advantage of the parametric model reduction technique for short-rate models. Since the reduced model of order $d = 10$ is an excellent approximation of the full model, we conclude that the full model valuations for the adaptively selected $\ell = 10$ parameter groups are enough, and the remaining of the parameter groups can be solved inexpensively using this reduced model.

References

1. M. Aichinger, A. Binder, *A Workout in Computational Finance*, 1st edn. (John Wiley and Sons Inc., West Sussex, UK, 2013)
2. A. Binder, O. Jadhav, V. Mehrmann, *Model order reduction for the simulation of parametric interest rate models in financial risk analysis*, J. Math. Industry 11, 8 (2021).

3. D. Brigo, F. Mercurio, *Interest Rate Models - Theory and Practice*, 1st edn. (Springer-Verlag, Berlin, 2006)
4. European Commission, in *Commission delegated regulation (EU) 1286/2014*, Off. J. EU. (Available via EUR-Lex, 2014), https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32014R1286
5. G. H. Golub, C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math. 14 (1970), 403–420.
6. J.C. Hull, A.D. White, *Numerical Procedures for Implementing Term Structure Models II Two-Factor Models*, J. Deriv. 2(2) (1994), 37–48.
7. K. Kunisch, S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math. 90 (2001), 117–148.
8. A. Paul-Dubois-Taine, D. Amsallem, *An adaptive and efficient greedy procedure for the optimal training of parametric reduced-order models*, Int. J. Numer. Meth. Engng, 102 (2015), 1262–1292.

A Low-Rank Extended Kalman Filter for Gas Pipeline Networks

Nadine Stahl and Nicole Marheineke

Abstract In this paper we deal with efficient state estimation of nonlinear partial differential-algebraic equations (PDAEs) with a low-rank Extended Kalman Filter version. We formulate a time-implicit Extended Kalman Filter and make use of model order reduction techniques to overcome the curse of dimensionality arising when discretizing PDAEs. As motivating example we look at a gas pipeline network described by the isothermal Euler equations. We compare our approach with other known low-rank Kalman filter variants.

1 Introduction

For control and optimization of gas transport, good knowledge of pressure and mass flux inside a pipe network is crucial. As usually only few measurements of these states are available, one relies on state estimation by combining a mathematical model with the actual measurements. When considering large networks, the models tend to have a high number of states and therefore become infeasible to be simulated or even optimized. Model order reduction gives a possibility to decrease the number of states while retaining a certain approximation quality.

Our gas pipeline network, modelled as a directed graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$, is described by the isothermal Euler equations on the space-time domain $[0, l^e] \times (0, T]$

$$a^e \partial_t p^e = -\partial_x q^e, \quad b^e \partial_t q^e = -\partial_x p^e - d^e q^e \frac{|q^e|}{p^e} \tag{1a}$$

on each pipe (edge) $e \in \mathcal{E}$ with length l^e and pipe parameters a^e, b^e, d^e. At the pipe junctions (nodes) $v \in \mathcal{V}$ we preserve mass flux and momentum through the Kirchhoff conditions, i.e., for δ_v^+, δ_v^- being the sets of all topologically ingoing and

N. Stahl (✉) · N. Marheineke
Trier University, Trier, Germany
e-mail: nadine.stahl@uni-trier.de; marheineke@uni-trier.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_24

outgoing pipes to a junction $v \in \mathcal{V}$, pressure and mass flux fullfil

$$\sum_{e\in\delta_v^-} q^e(l^e,t) = \sum_{e\in\delta_v^+} q^e(0,t), \quad p^e(l^e,t) = p^v(t), e \in \delta_v^+, \quad p^e(0,t) = p^v(t), e \in \delta_v^-. \tag{1b}$$

To close the system, we propose pressure values at every in- and outlet and assume consistent initial values for pressure and mass flux. The resulting system is well-posed, see [4].

In this paper we propose an efficient state estimator for (1) using the Extended Kalman Filter based on an implicit time-discretization and nonlinear model order reduction. We compare our results to, first, other Kalman Filter variants [1, 3] suitable for nonlinear models and, second, to state estimates done with a linearized reduced model [5].

2 Nonlinear Filtering

Based on [4], discretization in space of (1) with mixed finite elements yields a nonlinear system of the following form for $t \in (0, T]$:

$$E\dot{x}(t) = A(x(t)) + Bu(t), \tag{2}$$

with state $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^p$, $E \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,p}$ and $A\colon \mathbb{R}^n \to \mathbb{R}^n$. As this system is a DAE with a singular matrix E on the left side, we cannot use an explicit integration method for solving (2) in time. As in [5], we apply a general θ-scheme for time discretization. Dividing the time interval $[0, T]$ into K equidistant intervals of length τ, the time-discrete state of (2) at time $t_k := k\tau$ is denoted by x_k.

For the state estimation, we further introduce a state noise $\eta_k \sim \mathcal{N}(0, \tau^2 Q)$ in each time step to account for model inaccuracies, where $Q \in \mathbb{R}^{n,n}$ is a constant symmetric positive definite matrix. In the following, we set up an Extended Kalman Filter suitable for the above time discretization. Note that, the same calculations carry over to other Kalman Filter variants when using an implicit time-scheme.

2.1 *The Extended Kalman Filter*

The Kalman Filter [2] describes the distribution of the state $x(t)$ by its expectation value and its error covariance. For linear systems with Gaussian noise, these are sufficient to describe the distribution correctly. For nonlinear systems, we get a suboptimal filter, i.e. the distribution cannot be described by expectation value and covariance alone. The most common way to set up the filter is by linearizing the system in each time step in order to calculate the covariance matrices. This is then known as the Extended Kalman Filter (EKF) [6].

We follow [6], where an Extended Kalman Filter for explicit time-schemes is derived, to present an Extended Kalman Filter for implicit time-schemes. We introduce the predicted and corrected state estimates $x_{k|k-1}$ and $x_{k|k}$, where the first index corresponds to the time step at which the model is considered, whereas the second one corresponds to the time step up until measurements are included. Similarly, the predicted and corrected covariance matrices are denoted by $P_{k|k-1}$ and $P_{k|k}$.

The filter assumes that noisy measurements $y_k \in \mathbb{R}^q$ at time t_k are taken, i.e.

$$y_k = Hx_k + \nu_k, \tag{3}$$

where $H \in \mathbb{R}^{n,q}$ is an output matrix and $\nu_k \sim \mathcal{N}(0, R)$ is the measurement noise with constant covariance $R \in \mathbb{R}^{q,q}$.

Theorem 1 *The Extended Kalman Filter for a nonlinear system of form* (2) *and measurements derived from* (3) *with an underlying θ-scheme for time discretization has the following form:*

$$Ex_{k|k-1} - \theta\tau A(x_{k|k-1}) = Ex_{k-1|k-1} + (1-\theta)\tau A(x_{k-1|k-1}) + \tau B u_{\theta k}, \tag{4a}$$

$$P_{k|k-1} = \Phi_k P_{k-1|k-1}\Phi_k^T + \tau^2 Q, \tag{4b}$$

$$K_k = (P_{k|k-1}H^T)(HP_{k|k-1}H^T + R)^{-1}, \tag{4c}$$

$$x_{k|k} = x_{k|k-1} + K_k(y_k - Hx_{k|k-1}), \tag{4d}$$

$$P_{k|k} = (1 - HK_k)P_{k|k-1}, \tag{4e}$$

with

$$\Phi_k = (E - \theta\tau D_A(x_{k|k-1}))^{-1}(E + (1-\theta)\tau D_A(x_{k-1|k-1})), \tag{4f}$$

where D_A denotes the Jacobian of the nonlinearity A and with

$$u_{\theta k} := \theta u(t_k) + (1-\theta)u(t_{k-1}). \tag{4g}$$

Proof We start by deriving an equation to determine the expectation value of the state at time t_k, which will be denoted by x_k. Discretizing (2) in time and adding the state noise results in

$$Ex_k - \theta\tau A(x_k) = Ex_{k-1} + (1-\theta)\tau A(x_{k-1}) + \tau B u_{\theta k} + \eta_k. \tag{5}$$

We linearize the nonlinearity on the left side of equation (5) around $x_{k|k-1}$ and the right side around $x_{k-1|k-1}$, yielding

$$A(x_k) \approx A(x_{k|k-1}) + D_A(x_{k|k-1})(x_k - x_{k|k-1}) := D_A(x_{k|k-1})x_k + d_k,$$

$$A(x_{k-1}) \approx A(x_{k-1|k-1}) + D_A(x_{k-1|k-1})(x_{k-1} - x_{k-1|k-1}) := D_A(x_{k-1|k-1})x_{k-1} + e_k$$

respectively. Note that, d_k and e_k are deterministic and therefore later on, have only constant influence on the expectation value. The state x_k can then be represented as

$$\begin{aligned} x_k &= (E - \theta\tau D_A(x_{k|k-1}))^{-1} \\ &\quad [(E + (1-\theta)\tau D_A(x_{k-1|k-1}))x_{k-1} + \tau B u_{\theta k} + \eta_k + \tau((1-\theta)e_k - \theta d_k)]. \end{aligned}$$

Using this identity, exploiting the linearity of the expectation value and knowing that $\mathbb{E}[x_{k-1}] = x_{k-1|k-1}$ for the previous time step t_{k-1}, the predicted state estimate at time t_k is given as

$$\begin{aligned} x_{k|k-1} &= \mathbb{E}[x_k | \{y_i\}_{i=1}^{k-1}] \\ &= (E - \theta\tau D_A(x_{k|k-1}))^{-1} \\ &\quad [(E + (1-\theta)\tau D_A(x_{k-1|k-1}))x_{k-1|k-1} + \tau B u_{\theta k} + \tau((1-\theta)e_k - \theta d_k)]. \end{aligned}$$

Resubstituting e_k and d_k yields the proposed filter equation. Note that, in the equations above, we used the linearity of the expectation value to derive the iterating time-scheme for the state estimate.

Analogously we get the equation for the predicted covariance matrix

$$P_{k|k-1} := \mathbb{E}[(x_k - x_{k|k-1})(x_k - x_{k|k-1})^T | \{y_i\}_{i=1}^{k-1}].$$

The last three equations are the same as in the standard Kalman Filter and hence are not further treated here. □

2.2 *The Reduced Extended Kalman Filter*

System (2) is of high dimensionality and thus not suitable for an efficient state estimation. Using Model Order Reduction based on [4], which uses Proper Orthogonal Decomposition coupled with a complexity reduction for the nonlinear term, we end up with a system of the same form as (2) but of much smaller dimension $N \ll n$. Note that, the reduction technique from [4] preserves not only block structure for the reduced system, but also guarantees the reduced system to be stable. Additionally, mass and energy are preserved throughout simulation due to the port-Hamiltonian structure of the reduced system. Applying the Extended Kalman Filter (4) onto the reduced system then yields the Reduced Extended Kalman Filter (REKF).

2.3 Other Kalman Filter Variants

We now shortly review two other filtering methods based on the Kalman Filter, which we will use in Sect. 3.

The first one is the Ensemble Kalman Filter (EnKF) [1], which is based on an ensemble for which the filtering algorithm is executed. The predicted error covariance matrix is calculated in each time step as the covariance of the samples, whereas the Kalman gain and the corrected error covariance matrix are calculated according to the standard formulas. Each sample is then updated with the new measurement and the overall state estimate is given as the mean of the updated samples. This filter is combined with the reduced order model (REnKF).

The other variant is the Compressed State Kalman Filter (CSKF) [3]. This filter has similar ideas as we have, but instead of reducing the whole state dimension, here only the covariance matrices are assumed to have a low-rank representation. By help of this approximation, the highly costly calculations for the covariances and the Kalman gain are projected onto a low dimensional subspace, while the state simulation itself is kept in the full space. A prolongation of the Kalman gain is needed in every time-step to correctly update the state estimate with new measurements.

3 Numerical Example

As an academic example network we use the diamond network, see also Fig. 1, with the same pipe parameters and as in [5]. As inputs and measurements we choose

$$p^{v_1}(t) = \begin{cases} 2+t, & 0 \leq t < 1, \\ 3, & 1 < t < 5, \\ 3-0.1t, & 5 \leq t < 10, \\ 2.5, & t \geq 10, \end{cases} \qquad p^{v_2}(t) \equiv 2$$

and the fluxes at the two boundary nodes, i.e. at v_1 and v_2 for $t \in [0, 20]$, respectively, with $\tau = 0.02$ and $\theta = 0.51$.

We compare our nonlinear filtering approach with first linearizing the friction term in (1) and then applying the Kalman filter variants from Sect. 2 for linear models, see also [5] for more details on the linearized approach. Note that, in the case of using the Extended Kalman Filter for the nonlinear model, we have to linearize and evaluate Jacobians during the estimation step, which means, in every time step. In contrast, for the linearized system, the Jacobians are constant in time and have to be evaluated only once. We therefore expect latter method in to be significantly faster, whereas we assume our method presented here to be more accurate.

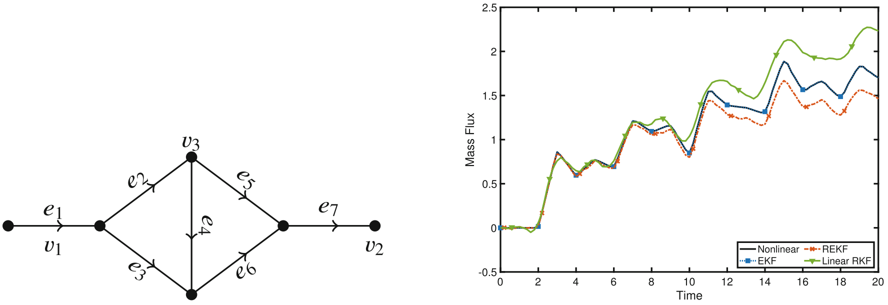


Fig. 1 Left: Diamond network topology. Right: Flux simulation and estimates at node v_3 with full nonlinear (EKF), reduced nonlinear (REKF) and reduced linear (Linear RKF) models

Table 1 Errors and Offline/Online times for different Kalman Filter variants

Method	Nonlinear filtering		Linear filtering		
	$\text{mean}_j \frac{\|\mathbb{E}[\mathbf{x}_j - \mathbf{x}_{j\|j}]\|_{L^2}}{\|\mathbb{E}[\mathbf{x}_j]\|_{L^2}}$	Online/s	$\text{mean}_j \frac{\|\mathbb{E}[\mathbf{x}_j - \mathbf{x}_{j\|j}]\|_{L^2}}{\|\mathbb{E}[\mathbf{x}_j]\|_{L^2}}$	Offline/s	Online/s
EKF	$1.1 \cdot 10^{-5}$	$2.5 \cdot 10^{3}$	$3.7 \cdot 10^{-2}$	$8.4 \cdot 10^{2}$	$5.2 \cdot 10^{1}$
REKF	$3.8 \cdot 10^{-2}$	$7.5 \cdot 10^{0}$	$6.0 \cdot 10^{-2}$	$0.2 \cdot 10^{0}$	$7.4 \cdot 10^{-2}$
CSKF	$3.8 \cdot 10^{-7}$	$1.0 \cdot 10^{2}$	$3.8 \cdot 10^{-2}$	$2.1 \cdot 10^{0}$	$1.5 \cdot 10^{0}$
REnKF	$3.9 \cdot 10^{-2}$	$6.6 \cdot 10^{2}$	$9.8 \cdot 10^{-2}$	–	$1.5 \cdot 10^{0}$

These expectations are verified in Table 1 and visualized in Fig. 1. Here, the mean relative errors with respect to the nonlinear full model are depicted using the nonlinear and linear state estimation methods. Clearly, the nonlinear methods behave better in terms of accuracy. Note that, the reduced order models in both cases were of approximation order 10^{-2} with respect to the respective full order model. The CSKF for the nonlinear case performs surprisingly well, but we observed that this algorithm tends to avoid measurement updates due to some cancelations which lead to trivial Kalman gains. Also for some projection matrices resulted from model order reduction, the CSKF was unstable. Concerning the computational times, using the linear Kalman filter variants easily outperforms using the nonlinear ones, as we expected. Another major disadvantage of the Extended Kalman Filter is, that we cannot divide the algorithm into an offline and an online phase, whereas in the linear case, we can pre-compute the covariance matrices and Kalman gains independent of the actual measurements and estimates in the offline phase.

Summarizing, our Extended Kalman Filter using an implicit time-scheme was very computational demanding even when using a reduced order model. In comparison to recent investigations on using a linearized reduced order model for the estimation instead, we clearly were outperformed in terms of computational time. Although our approximation errors are slightly better, using an implicit time scheme for a nonlinear state estimation seems to add a lot more computational effort while the gain in accuracy is only minor.

References

1. Houtekamer, P.L., Mitchell, Herschel L.: Ensemble Kalman Filtering. Q. J. R. Meteorolog. Soc. **131** (2006), 3269–3289.
2. Kalman R.E: A new approach to linear filtering and prediction problems. J. Fluids Eng. **82** (1960), 35–45
3. Li, J.Y., Kokkinaki, A., Ghorbanidehno, H., Darve, E.F., Kitanidis, P.K.: The compressed state Kalman filter for nonlinear state estimation: Application to large-scale reservoir monitoring. Water Resour. Res. **51** (2015), 9942–9963.
4. Liljegren-Sailer, B., Marheineke, N.: Port-Hamiltonian approximation of a nonlinear flow problem. Part I: Space approximation ansatz. arXiv:2009.11216
5. Stahl N., Marheineke N.: Efficient state estimation for gas pipeline networks via low-rank approximations. arXiv:2007.15988
6. Yu, B., Shenoy, K., Sahani, M.: Derivation of Extended Kalman Filtering and Smoothing Equations. Stanford University, Tech. Rep. (2004)

An Analysis of Connectivity Between Dengue Cases and Climate Factors in Sri Lanka Based on Field Data

Hasitha Erandi, Karunia Putra Wijaya, Naleen Ganegoda, and Thomas Goetz

Abstract Dengue is the most critical mosquito-borne viral disease that has rapidly spread within recent years. Understanding of the seasonal pattern of dengue cases and relationship with climate data could be useful in deciding control mechanisms. In this study, monthly dengue cases, average rainfall data, average temperature data and relative humidity data of each province in Sri Lanka from 2010 to 2019 have been analyzed to identify the periodic pattern and the delayed effect of climate factors on dengue cases. First, we have used the Fast Fourier Transform (FFT) to identify the periodic patterns of dengue cases and climate data. Next, we have used the Pearson's correlation coefficient to find the time delay between climate data and dengue cases. The results reflected that out of nine provinces, dengue cases in Western, Central, Southern, North Western and North Central provinces are influenced by both monsoon seasons. Moreover, in Western, Southern, North Western, Sabaragamuwa, Northern and Eastern provinces, periodic pattern of dengue cases follows the periodic pattern of the rainfall data with two months time delay. The delayed effect of average temperature on dengue cases is three months and that of relative humidity is one/two months. These results could be used in health planning during the outbreaks.

H. Erandi (✉)
Department of Mathematics, University of Colombo, Colombo, Sri Lanka

K. P. Wijaya · T. Goetz
Mathematical Institute, University of Koblenz, Koblenz, Germany
e-mail: karuniaputra@uni-koblenz.de; goetz@uni-koblenz.de

N. Ganegoda
Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka
e-mail: naleen@sjp.ac.lk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_25

1 Introduction

Dengue is a viral infection which is transmitted by bites of infected female *Aedes Aegypti* or *Ades Albopictus* mosquitoes. The number of reported dengue cases has been rapidly increased during the past few decades and it has become one of the major public health issues in tropical and subtropical regions in the world [1]. The disease is endemic in more than 100 countries and 1.3 million cases are reported annually with 2.5% death rate. The first dengue case in Sri Lanka was reported in 1962 and now it has gained the endemic stage. The country experienced its first outbreak during 1965–1966 [2]. Since then, several dengue outbreaks occurred and the worst was reported in 2017 with total 186,101 cases. This is a 3.4 fold higher than the total reported cases from 2016.

Dengue is caused by four virus serotypes, namely DENV-1, DENV-2, DENV-3 and DENV-4 which are serologically related and infection of one serotype does not provide cross immunity against the other three serotypes [2]. Thus, a person can be infected with multiple serotypes during their lifetime and a person with repeated infection has a risk of developing a severe morbidity level known as dengue hemorrhagic fever (DHF). Since specific drugs are not currently available and developing a vaccine for the disease has been trialing out, the main prevention strategy is vector control [3]. It is evident that vector control is also a challenge as mosquitoes become resistant and adaptable to the commonly used insecticides and control strategies [4]. Mosquito density is a highly sensitive factor which heavily depends upon climatic changes such as rainfall, humidity and temperature.

Several previous works have examined the influence of climate factors on dengue cases in different regions of Sri Lanka [5–8]. Further, the study in [5] identified the positive correlation between dengue cases and rainfall data in Gampaha district. However, the study in [6] concluded that weekly rainfall slightly influence dengue cases in Colombo and Anuradhapura and no influence in Ratnapura. Some studies concluded that reported dengue cases have strong correlation with rainfall in Colombo with different time lags [7, 8]. However, almost all studies have been focused on one or few regions in the country and there is no study for all provinces in Sri Lanka. Hence, detection of distribution pattern and relationship of dengue data with the climate data in each province is important in decision making to allocate resources for the disease prevention in different climate seasons.

In this study, first we analyze evolution of the dengue cases in Sri Lanka from year 2010 to 2019 and inter-provincial cross-correlation among reported dengue cases. Then we examine the correlation of dengue cases with climate factors to examine the delay effect on dengue. Finally, we determine the range for precipitation data which gives the highest correlation with dengue case.

2 Material and Methods

Sri Lanka is a tropical country located in the Indian Ocean, southwest of the Bay of Bengal, between 5°55′ to 9°51′ North latitude and 79°42′ to 81°53′ East longitude. The country is influenced by two monsoon seasons Northeast from December to January and Southwest from May to September and the disease is more intense every year within or soon after the monsoon seasons. Moreover, temperature of Sri Lanka varies from 17° to 35° which is ideal for vector biology. Based on the administrative system, the country is divided into 9 provinces. For each province, we use recorded monthly dengue cases from 2010 January to 2019 December obtained from Epidemiology unit, Ministry of Health, Sri Lanka and average rainfall, average temperature and relative humidity data from 2010 January to 2019 December obtained from Meteorological Department, Sri Lanka.

To identify the periodic patterns of the dengue cases and climate data, we need to extract frequency domain features from the data. Therefore, we use fast Fourier transformation to convert time domain data into frequency domain. To analyze the correlation between dengue and climate data we use Person correlation formula. We implement computational tools in *MATLAB* for this paper.

3 Interrelationship Between Provinces

To have an idea about dengue transmission pattern in each province we consider the monthly dengue cases, average rainfall data, average temperature data and relative humidity data from January 2010 to December 2019. From Fig. 1, it can be observed that distribution pattern of dengue cases is different from province to province. Western province has two peaks in each year and Northern province has a peak at the beginning of each year while the Uva province has no regular pattern. These factors motivate and thrive us to calculate the cross-correlation of dengue cases between provinces to identify the distribution pattern of dengue in Sri Lanka.

Though there are many methods to compute the correlation coefficient, Pearson correlation coefficient is the standard method to compute the strength of the linear relationship between two variables [3]. Hence, the cross-correlation of dengue cases between provinces have been calculated using Pearson correlation formula. The results reflect that, Sabaragamuwa and Uva provinces are influenced only by Southwest monsoon and Northern and Eastern provinces are influenced only by Northeast monsoon. Western, Central, Southern, North Western and North Central provinces are influenced by both Southwest and Northeast monsoons. Also, we have observed that dengue cases in Northern province follows the periodic pattern of dengue cases in Eastern province with one month time lag. Since the epidemiological data usually consists of multiple periodic components, next task is to identify the periodic patterns of dengue cases and climate data.

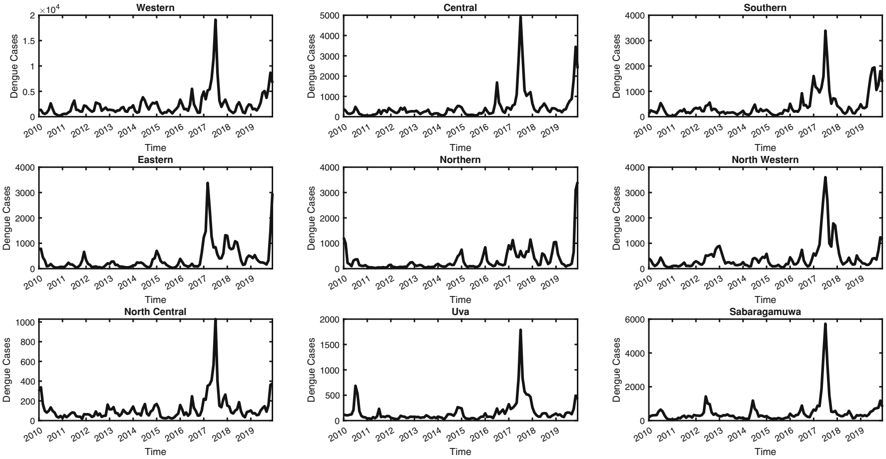


Fig. 1 Reported monthly dengue cases in each province from January 2010 to December 2019

4 Periodic Structure

To identify the periodic pattern, Fourier spectrum of dengue cases and climate data from 2010 January to 2019 December has been calculated for each province [3]. Table 1 represents the summary of the results for periodic pattern. From Table 1 it can be observed that the dengue cases in Western, Central, Southern, North Western and North Central provinces exhibit 6 months periodic patterns. In addition, Northern, Eastern, Sabaragamuwa and Uva provinces exhibit an annual periodic pattern. Further, notice that there is a Fourier amplitude related to 6 months for each Northern, Eastern, Sabaragamuwa and Uva provinces. Moreover, it can be observed that the rainfall data in Western, Southern, Central, North Western, Sabaragamuwa and Uva provinces show 6 months periodic pattern while the rainfall data in North Central, Northern and Eastern provinces show an annual periodic pattern. Moreover, we have noticed that dengue cases in all provinces have a three year cycle. Results for the periodic patterns of reported dengue cases, indicate that dengue cases have a relationship with rainfall. Therefore, it is worthwhile to analyze the cross-correlation of reported dengue cases with climate data.

5 Correlation with Climate Data

Factors like incubation period motivate us to measure the correlation between climate data and dengue cases with time delay. To calculate the time delay, the Pearson correlation formula [3] has been used. Maximal correlation coefficients

Table 1 Period of monthly dengue cases and climate data extracted using fast Fourier transform in all the provinces, Sri Lanka. (Data: 2010–2019)

Province	Dengue	Rainfall	Temperature	Humidity
Western	6	6	12	12
Southern	6	6	12	6
Central	6	6	12	12
North Western	6	6	12	6
North Central	6	12	12	12
Northern	12	12	12	12
Eastern	12	12	12	12
Sabaragamuwa	12	6	12	6
Uva	12	6	12	12

Table 2 Maximal correlation coefficients between dengue cases and climate data subject to different time lags in all the provinces, Sri Lanka. (Data: 2010–2019, Time lag: 0–6)

Province	Rainfall	Temperature	Humidity
Western	0.2363 (2)	0.3181 (3)	0.2000 (1)
Southern	0.2460 (2)	0.3402 (3)	−0.0650 (2)
Central	−0.0605 (2)	0.0187 (0)	0.1652 (1)
North Western	0.0978 (2)	0.2197 (3)	0.1524 (1)
North Central	0.1776 (2)	0.1705 (3)	0.0708 (2)
Northern	0.3593 (2)	0.3419 (6)	0.3876 (1)
Eastern	0.1835 (2)	0.2744 (6)	0.2860 (2)
Sabaragamuwa	0.2625 (2)	0.2986 (3)	0.1082 (1)
Uva	0.0187 (2)	0.3850 (2)	−0.0513 (2)

between dengue cases and climate data subject to different time lags in all the provinces are depicted in Table 2.

From Table 2 it can be observed that the highest correlation between dengue cases and rainfall data occurs with a 2-month delay for each province. For temperature data, most of the province delay time is 3-months and the delayed effect of relative humidity on dengue cases is one/two months.

However, heavy rainfall can potentially flush away larvae or pupae or the immature stage of mosquitoes and increase the mortality rate of adult mosquitoes. Therefore, finding minimum and maximum cutoff points for rainfall data which give highest correlation value with dengue cases is important for decision makers to predict the number of dengue cases in upcoming monsoon seasons. For each province, we have checked interval of confidence, highest correlation value and time lag by increasing minimum cutoff value and decreasing maximum cutoff value. Then the best match cutoff values for rainfall data have been found with more than 55% confidence interval for 5 provinces. Table 3 represents the minimum and maximum cutoff values with new time lag.

Table 3 Minimum and maximum cutoff values with new time lag

Province	Minimum cutoff (mm)	Maximum cutoff (mm)	New time lag (months)
Western	82.2	724.8	3
Northern	17	425.8	2
Eastern	45.1	883.8	3
North West	0	568.25	3
North Central	45	600	3

6 Conclusion

The main purpose of this study was to analyze the pattern of monthly dengue cases and its relationship with climate data of each province in Sri Lanka. We observed that dengue cases in 5 provinces; Western, Central, Southern,North Western and North Central are influenced by both Northeast and Southwest monsoon while other 4 provinces influenced by only one monsoon season. In Western, Southern, North Western, Sabaragamuwa, Northern and Eastern provinces out of 9 provinces, periodic pattern of dengue cases follows the periodic pattern of the rainfall data and the delayed effect of rainfall data on dengue cases is two months. Moreover, the average temperature data of all provinces have annual periodic pattern and the delayed effect on dengue cases is three months for most of the provinces. The delayed effect of relative humidity on dengue cases is one/two months. Since we have used 10 years data for the study, these results comprehensively extract long term variation too.

Acknowledgments This work was partially supported by National Science Foundation, Sri Lanka. Grant number RPHS/2016/D/05.

References

1. Messina, J.P., Brady, O.J., Scott, T.W., Zou, C., Pigott, D.M., Duda, K.A., Bhatt, S., Katzelnick, L., Howes, R.E., Battle, K.E. and Simmons, C.P. Global spread of dengue virus types: mapping the 70 year history. Trends in Microbiology, 22(3), 138–146, 2014.
2. Sirisena, P.D.N.N. and Noordeen, F., Evolution of dengue in Sri Lanka – changes in the virus, vector, and climate. International Journal of Infectious Diseases, 19, 6–12, 2014.
3. Erandi, K.K.W.H., Perera, S.S.N. and Mahasinghe, A.C., Analysis and forecast of dengue incidence in urban Colombo, Sri Lanka, Theoretical Biology and Medical Modelling, 18(1), 1–19, 2021.
4. Bhatia, R., Dash, A.P. and Sunyoto, T. Changing epidemiology of dengue in South-East Asia. WHO South-East Asia Journal of Public Health, 2(1), 23, 2013.
5. Arunachalam, N., Tana, S., Espino, F., Kittayapong, P., Abeyewickrem, W., Wai, K.T., Tyagi, B.K., Kroeger, A., Sommerfeld, J. and Petzold, M., Eco-bio-social determinants of dengue vector breeding: a multicountry study in urban and periurban Asia. Bulletin of the World Health Organization, 88, 173–184, 2010.

6. Goto, K., Kumarendran, B., Mettananda, S., Gunasekara, D., Fujii, Y. and Kaneko, S., Analysis of effects of meteorological factors on dengue incidence in Sri Lanka using time series data. PloS One, 8(5), e63717, 2013.
7. Pathirana, S., Kawabata, M. and Goonatilake, R., Study of potential risk of dengue disease outbreak in Sri Lanka using GIS and statistical modelling. Journal of Rural and Tropical Public Health, 8, 8–17, 2009.
8. Wickramaarachchi, W.P.T.M. and Perera, S.S.N., A mathematical model with control to analyse the dynamics of dengue disease transmission in urban Colombo. Journal of National Science Foundation Sri Lanka, 46(1), 41–49, 2018.

Cellular Nonlinear Computing on the Edge of Chaos

Angela Slavova

Abstract In this paper we study Cellular Nonlinear Networks (CNN) working on the edge of chaos. First we present a reaction-diffusion CNN model. Edge of chaos regime is determined for this model based on local activity theory. Numerical simulation is presented in order to illustrate the obtained theoretical results.

1 Introduction

Cellular Nonlinear Networks (CNN) present a new class of information processing systems which shows important potential applications. The concept of CNN is based on some aspects of neurobiology and adapted to integrated circuits. CNN are defined as spatial arrangements of locally coupled dynamical systems, referred to as cells. The CNN dynamics are determined by a dynamic law of an isolated cell, by the coupling laws between the cells and by boundary and initial conditions. The cell coupling is confined to the local neighborhood of a cell within a defined sphere of influence. The dynamic law and the coupling laws of a cell are often combined and described by nonlinear ordinary differential- or difference equations (ODE), respectively, referred to as the state equations of cells. Thus a CNN is given by a system of coupled ODE with a very compact representation in the case of translation invariant state equations. Despite of having a compact representation, CNN can show complex dynamics like chaotic behavior, self-organization, and pattern formation or nonlinear oscillation and wave propagation. Furthermore, Reaction-Diffusion Cellular Nonlinear/Nanoscale Networks (RD-CNN) have been applied for modeling complex systems [1, 4–6]. These networks are not representing a paradigm for complexity only but also establishing novel approaches to information processing by the dynamics of nonlinear complex systems.

A. Slavova (✉)
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
e-mail: slavova@math.bas.bg

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_26

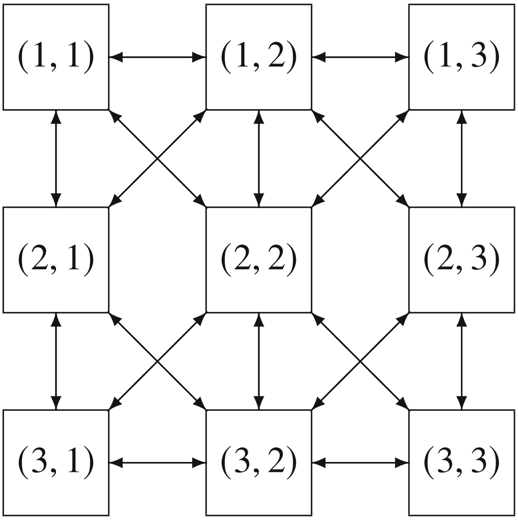


Fig. 1 CNN with 3 × 3 neighbourhood system

Let us consider a two-dimensional grid with 3 × 3 neighborhood system as it is shown in Fig. 1.

2 Reaction-Diffusion CNN

In this chapter we shall present the derivation of the CNN implementations via spatial discretization, which suggests a methodology for converting a PDE to CNN templates and vice versa. The CNN solution of a PDE has four basic properties—it is

1. continuous in time;
2. continuous and bounded in value;
3. continuous in interaction parameters;
4. discrete in space.

Reaction-diffusion CNN are described mathematically by a discretized version of the well-known system of nonlinear PDEs—reaction-diffusion equations [7]:

$$\frac{\partial u}{\partial t} = f(u) + D\nabla^2 u, \tag{1}$$

where $u \in \mathbb{R}^m$, $f \in \mathbf{R}^m$, D is a $m \times m$ diagonal matrix whose diagonal elements D_i are called the diffusion coefficients, and

$$\nabla^2 u_i = \frac{\partial^2 u_i}{\partial x^2} + \frac{\partial^2 u_i}{\partial y^2}, \quad i = 1, 2, \ldots, m, \tag{2}$$

is the Laplacian operator in $\mathbb{R}^2$.

There are several ways to approximate the Laplacian operator (2) in discrete space by a CNN synaptic law with an appropriate A-template [1, 5]. For example we can have:

(a) one-dimensional discretized Laplacian template:

$$A_1 : (1, -2, 1); \tag{3}$$

(b) two-dimensional discretized Laplacian template:

$$A_2 : \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{4}$$

3 Edge of Chaos Regime

In this section we shall apply theory of local activity [2, 3] in order to study the dynamics of reaction-diffusion CNN. The theory which will be presented below offers both constructive analytical and numerical method for obtaining local activity of reaction-diffusion CNN. It is known [3] that for reaction-diffusion CNN, one can determine the domain of the cell parameters such that the cells are locally active, and thus potentially capable of exhibiting complexity. This precisely defined parameter domain is called edge of chaos. We shall present below constructive and explicit mathematical inequalities for identifying the region in the parameter space where complexity phenomena may emerge. By restricting the cell parameter space to the local activity domain we can achieve a major reduction in the computing time required by the parameter search algorithms. In this way there is a possibility of exploiting and controlling chaos for future scientific and engineering applications [2].

The model of hysteresis CNN which we shall study is the following:

$$\frac{du_i}{dt} = -u_i - 2h(u_i) + d(\sigma_i^1 u_0 - u_i), \tag{5}$$

$$\frac{du_0}{dt} = -u_0 - 2h(u_0).$$

In this paper we develop the following constructive algorithm for determining the edge of chaos domain:

1. Find the equilibrium points E_j, $j = 1, 2$ of hysteresis CNN model (5).
3. Calculate the cell coefficients of the Jacobian matrix about each system equilibrium point E_j, $i = 1, 2$.

4. Calculate the trace $Tr(E_j)$ and the determinant $\Delta(E_j)$ of the Jacobian matrix for each equilibrium point.
5. Define stable and locally active region $SLAR(E_j)$ for the equilibrium points:

Definition 1 Stable and locally active region $SLAR(E_j)$ at the equilibrium point E_j for the hysteresis CNN model is such that $Tr(E_j) < 0$ and $\Delta(E_j) > 0$.

6. Determine Edge of chaos region.

In the literature, the so-called edge of chaos (EC) means a region in the parameter space of a dynamical system where complex phenomena and information processing can emerge. We shall try to define more precisely this phenomena till now known only via empirical examples. Moreover, we shall present an algorithm for determining the edge of chaos for hysteresis CNN.

Definition 2 A hysteresis CNN (5) is said to be operating on the edge of chaos EC iff there is at least one equilibrium point E_j, $i = 1, 2$ which is both locally active and stable.

The following Theorem hold:

Theorem 1 *Hysteresis CNN model with dynamic memory synapses (5) is operating in the edge of chaos region iff* $d(2\sigma_i^1 - 1) < 1$ *and* $d(2\sigma_i^1 + 1) > -2$. *For this parameter value there is at least one equilibrium point which belongs to* $SLAR(E_j)$.

Proof We first define the equilibrium points of CNN model (5)—$E_1 = (\frac{-2-2d\sigma_i^1}{1+d}, -2)$, $E_2 = (\frac{2+2d\sigma_i^1}{1+d}, 2)$. We define the 4 cell coefficients of the Jacobian matrix for (5) and $Tr(E_j)$, $\Delta(E_j)$. Then according to definition 2 we find the inequalities of the parameter set for which we have at least one stable and locally active equilibrium points: $d(2\sigma_i^1 - 1) < 1$ and $d(2\sigma_i^1 + 1) > -2$. □

Numerical simulations in Fig. 2 show the following edge of chaos domain:

4 Conclusion

In this paper we study the model (5) of reaction diffusion CNN, which is actually a system of discrete ordinary differential equations. We determine the edge of chaos regime for this model. Based on the local activity theory we develop constructive algorithm for obtaining the edge of chaos regime. We provide numerical simulations for the following parameter set: $d(2\sigma_i^1) < 1$, $d(2\sigma_i^1 + 1) > -1$ and taking different values of the constant d and the binary pattern σ^1 in the case of 4 cells.

Acknowledgments The first author acknowledges as well the provided access to the e-infrastructure of the Centre for Advanced Computing and Data Processing, with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014–2020) and co-financed by the European Union through the European structural and Investment funds.

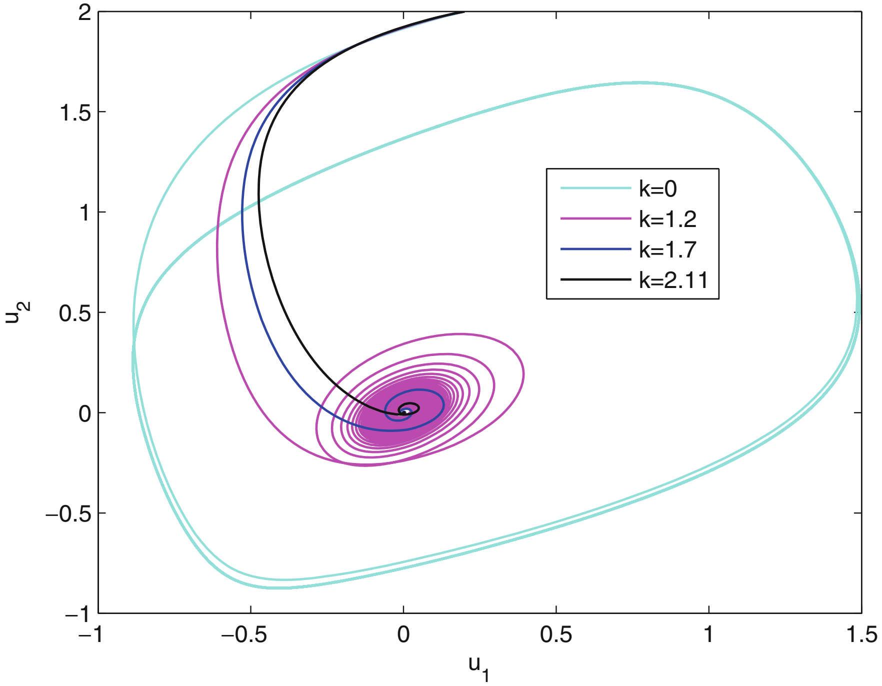


Fig. 2 Simulation of edge of chaos region

References

1. Chua, L.O., Yang, L.: Cellular Neural Network: Theory and Applications. IEEE Trans. CAS. vol. 35, p. 1257, (1988)
2. Chua, L.O.: Local Activity is the origin of complexity, Int. J. Bifurcation and Chaos, vol. 15, No. 11, pp. 3435–3456, (2005)
3. Mainzer, K., Chua, L.O.: Local Activity Principle: The Cause of Complexity and Symmetry Breaking, London: Imperial College Press, (2013)
4. Manganaro, G., Arena, P., Fortuna, L.: Cellular Neural Networks: Chaos, Complexity and VLSI Processing, Springer Verlag, (1999)
5. Slavova, A.: Cellular Neural Networks: Dynamics and Modeling, Kluwer Academic Publishers, 2003.
6. Slavova, A.: Tetzlaff, R.: Edge of chaos in reaction diffusion CNN model. Open Mathematics, vol.15, issue 1, (2017), https://doi.org/10.1515/math-2017-0002
7. Vidyasagar, M.: Nonlinear systems analysis, 2nd ed., Philadelphia:Society for Industrial and Applied Mathematics, (2002)

Efficient Yield Optimization with Limited Gradient Information

Mona Fuhrländer and Sebastian Schöps

Abstract An efficient strategy for yield optimization with uncertain and deterministic optimization variables is presented. The gradient based adaptive Newton-Monte Carlo method is modified, such that it can handle variables with (uncertain parameters) and without (deterministic parameters) analytical gradient information. This mixed strategy is numerically compared to derivative free approaches.

1 Introduction

In mass production one often has to deal with uncertainties due to manufacturing imperfections, which lead to deviations in the specified design parameters, i.e., geometry or material parameters, of the manufactured device. These deviations in the design parameters, may lead to deviations in the performance quantities, such that predefined performance requirements are not fulfilled. Thus, the device is useless. This is of course a waste of time, money and resources – and should be avoided.

In order to quantify the uncertainty, we consider the yield as the so-called *probability of success*. It is defined as the percentage of realizations in a manufacturing process, which fulfills all performance requirements, taking into account manufacturing uncertainties [6]. The yield can be estimated e.g. by a Monte Carlo (MC) analysis [7, Chap. 5]. In this work we will focus on the optimization procedure, i.e., the maximization of the yield in order to reduce the negative impact of uncertainty. In [3, 6] gradient based optimization algorithms have been proposed, assuming that gradients are available in analytical form. But this is only the case under some suitable conditions. In the following we present a strategy for efficient

M. Fuhrländer (✉) · S. Schöps
Computational Electromagnetics Group (CEM) and Centre for Computational Engineering (CCE), TU Darmstadt, Darmstadt, Germany
e-mail: mona.fuhrlaender@tu-darmstadt.de; sebastian.schoeps@tu-darmstadt.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_27

yield optimization under the assumption that only some of the partial derivatives are available.

2 Definition of the Yield

We define three kinds of parameters: uncertain design parameters, deterministic design parameters and range parameters. The uncertain parameters $\mathbf{p}$ are modeled as normal distributed random variables, i.e., $\mathbf{p} \sim \mathcal{N}(\overline{\mathbf{p}}, \mathbf{\Sigma})$, with

$$\mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}},\mathbf{\Sigma})}(\mathbf{p}) = \det(2\pi\mathbf{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{p}-\overline{\mathbf{p}})^\top \mathbf{\Sigma}^{-1}(\mathbf{p}-\overline{\mathbf{p}})\right), \tag{1}$$

where $\overline{\mathbf{p}} \in \mathbb{R}^{n_\mathbf{p}}$ indicates the mean value, $\mathbf{\Sigma} \in \mathbb{R}^{n_\mathbf{p}\times n_\mathbf{p}}$ the covariance matrix and $\mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}},\mathbf{\Sigma})}$ the corresponding probability density function (pdf). The deterministic parameters are given by $\mathbf{d} \in \mathbb{R}^{n_\mathbf{d}}$. The range parameter is denoted by $r \in T_r \subset \mathbb{R}$ and describes the environment in which the requirements have to be fulfilled.

Let $Q\colon \mathbb{R}^{n_\mathbf{p}+n_\mathbf{d}+1} \to \mathbb{R}$ be a quantity of interest (QoI) and $c \in \mathbb{R}$. We define the performance feature specifications (pfs) as

$$Q_r(\mathbf{p},\mathbf{d}) \leq c \quad \forall r \in T_r, \tag{2}$$

which can be easily extended to a vector-valued formulation in case of several requirements. Then the safe domain is the set of all parameter combinations fulfilling the pfs, and it depends on the current value of $\mathbf{d}$, i.e.,

$$\Omega_\mathbf{d} = \{\mathbf{p} : Q_r(\mathbf{p},\mathbf{d}) \leq c \;\; \forall r \in T_r\}. \tag{3}$$

The yield Y defines the percentage of realizations in a manufacturing process, which fulfill the pfs. Following [6] it is given by

$$Y(\overline{\mathbf{p}},\mathbf{d}) := \mathbb{E}[\mathbf{1}_{\Omega_\mathbf{d}}(\mathbf{p})] := \int_{\mathbb{R}^{n_\mathbf{p}}} \mathbf{1}_{\Omega_\mathbf{d}}(\mathbf{p})\,\mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}},\mathbf{\Sigma})}(\mathbf{p})\,\mathrm{d}\mathbf{p}, \tag{4}$$

where $\mathbb{E}$ denotes the expected value and $\mathbf{1}_{\Omega_\mathbf{d}}$ the indicator function with value 1 if the parameter $\mathbf{p}$ lies inside the safe domain and 0 otherwise.

A straightforward approach for yield estimation is MC analysis [7, Chap. 5]. There, a large set of sample points $\mathbf{p}^{(1)},\ldots,\mathbf{p}^{(N_\mathrm{MC})}$ of the uncertain parameter is randomly generated according to its pdf. Then the yield can be estimated by

$$Y(\overline{\mathbf{p}},\mathbf{d}) \approx Y_\mathrm{MC}(\overline{\mathbf{p}},\mathbf{d}) = \frac{1}{N_\mathrm{MC}} \sum_{i=1}^{N_\mathrm{MC}} \mathbf{1}_{\Omega_\mathbf{d}}(\mathbf{p}^{(i)}). \tag{5}$$

In computational engineering, the QoI often involves solving partial differential equations numerically, e.g., with a finite element method (FEM). Hence, it is computationally very expensive or even prohibitive to evaluate the QoI for the many sample points required in a MC analysis. For that reason, there is research on efficient yield estimation, using e.g. importance sampling [5], surrogate modeling [1, 10] or hybrid approaches [3, 4, 8]. These hybrid approaches combine classic MC with surrogate methods, e.g. Gaussian process regression (GPR), cf. [4]. Since this work focuses on the optimization process, we will not go into the details here.

3 Yield Optimization

We aim to maximize the yield by modifying the design, i.e.,

$$\max_{\overline{\mathbf{p}},\mathbf{d}} Y(\overline{\mathbf{p}},\mathbf{d}). \tag{6}$$

Let us assume that we have only uncertain design parameters as optimization variables and all of them are Gaussian distributed, i.e., $\max_{\overline{\mathbf{p}}} Y(\overline{\mathbf{p}},\mathbf{d})$. Then there exist closed form solutions of gradient and Hessian, cf. [6],

$$\nabla_{\overline{\mathbf{p}}} Y(\overline{\mathbf{p}},\mathbf{d}) = \int_{\mathbb{R}^{n_\mathbf{p}}} \mathbf{1}_{\Omega_\mathbf{d}}(\mathbf{p})\, \nabla_{\overline{\mathbf{p}}} \mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}},\mathbf{\Sigma})}(\mathbf{p})\, \mathrm{d}\mathbf{p}, \tag{7}$$

$$\nabla^2_{\overline{\mathbf{p}}} Y(\overline{\mathbf{p}},\mathbf{d}) = \int_{\mathbb{R}^{n_\mathbf{p}}} \mathbf{1}_{\Omega_\mathbf{d}}(\mathbf{p})\, \nabla^2_{\overline{\mathbf{p}}} \mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}},\mathbf{\Sigma})}(\mathbf{p})\, \mathrm{d}\mathbf{p}, \tag{8}$$

since the optimization variable $\overline{\mathbf{p}}$ only appears in the pdf and this is just an exponential function in case of Gaussian distribution. The MC estimators of the gradient and the Hessian are given by

$$\nabla_{\overline{\mathbf{p}}} Y_{\mathrm{MC}}(\overline{\mathbf{p}},\mathbf{d}) = Y_{\mathrm{MC}}(\overline{\mathbf{p}},\mathbf{d})\,\Sigma^{-1}\left(\overline{\mathbf{p}}_{\Omega_\mathbf{d}} - \overline{\mathbf{p}}\right), \tag{9}$$

$$\nabla^2_{\overline{\mathbf{p}}} Y_{\mathrm{MC}}(\overline{\mathbf{p}},\mathbf{d}) = Y_{\mathrm{MC}}(\overline{\mathbf{p}},\mathbf{d})\,\Sigma^{-1}\left(\mathbf{\Sigma}_{\Omega_\mathbf{d}} + \left(\overline{\mathbf{p}}_{\Omega_\mathbf{d}} - \overline{\mathbf{p}}\right)\left(\overline{\mathbf{p}}_{\Omega_\mathbf{d}} - \overline{\mathbf{p}}\right)^\top - \mathbf{\Sigma}\right)\Sigma^{-1}, \tag{10}$$

where $\overline{\mathbf{p}}_{\Omega_\mathbf{d}}$ indicates the mean value of all MC sample points lying inside the safe domain and $\mathbf{\Sigma}_{\Omega_\mathbf{d}}$ the corresponding covariance matrix. The detailed derivation can be found in [6]. Using (9) and (10), once the yield is estimated with MC, the derivatives are obtained without any additional computational effort. This allows to use a gradient based optimization solver, e.g. a globalized Newton method, cf. [11].

In [3] an adaptive Newton-MC method is proposed, which is an efficient modification of the globalized Newton method using the standard deviation of the MC estimation

$$\sigma_{\mathrm{MC}}(\overline{\mathbf{p}}, \mathbf{d}) = \sqrt{\frac{Y_{\mathrm{MC}}(\overline{\mathbf{p}}, \mathbf{d})(1 - Y_{\mathrm{MC}}(\overline{\mathbf{p}}, \mathbf{d}))}{N_{\mathrm{MC}}}} \tag{11}$$

as an error indicator for an adaptive sample size increase. For details we refer to [3].

Back to problem (6), we have uncertain *and* deterministic optimization variables. Since **d** appears in the indicator function, we cannot calculate the gradient of the yield with respect to **d**, given by

$$\nabla_{\mathbf{d}} Y(\overline{\mathbf{p}}, \mathbf{d}) = \int_{\mathbb{R}^{n_{\mathbf{p}}}} \nabla_{\mathbf{d}} \mathbf{1}_{\Omega_{\mathbf{d}}}(\mathbf{p})\, \mathrm{pdf}_{\mathcal{N}(\overline{\mathbf{p}}, \boldsymbol{\Sigma})}(\mathbf{p})\, \mathrm{d}\mathbf{p}, \tag{12}$$

analytically. Same holds for the Hessian. In order to still use the globalized Newton method or the adaptive Newton-MC, we propose a mixed strategy. We calculate the gradient with respect to **d** with finite differences. But we still use the analytical form for the derivative with respect to **p**. So we have

$$\nabla_{\overline{\mathbf{p}}, \mathbf{d}} Y(\overline{\mathbf{p}}, \mathbf{d}) = \left(\nabla_{\overline{\mathbf{p}}} Y(\overline{\mathbf{p}}, \mathbf{d}), \nabla_{\mathbf{d}} Y(\overline{\mathbf{p}}, \mathbf{d})\right)^{\top}, \tag{13}$$

where the first part is calculated with (7) and the second part with finite differences. A well-known formula to approximate Hessians is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update [11], given by

$$\mathbf{H}_{k+1}^{\mathrm{BFGS}} = \mathbf{H}_k + \frac{\mathbf{g}_k \mathbf{g}_k^{\top}}{\mathbf{g}_k^{\top} \mathbf{x}_k} - \frac{\mathbf{H}_k \mathbf{x}_k (\mathbf{H}_k \mathbf{x}_k)^{\top}}{\mathbf{x}_k^{\top} \mathbf{H}_k \mathbf{x}_k}, \tag{14}$$

where $\mathbf{H}_k$ is the Hessian from the last iterate, $\mathbf{g}_k$ the difference between the current and the last gradient and $\mathbf{x}_k$ the difference between the current and the last solution. Since the part of the Hessian belonging to the uncertain parameter can be calculated analytically by (8), we introduce the mixed BFGS Hessian

$$\mathbf{H}_{\mathrm{mix}}^{\mathrm{BFGS}} := \left(\begin{array}{c|c} \nabla_{\overline{\mathbf{p}}}^2 Y(\overline{\mathbf{p}}, \mathbf{d}) & \mathbf{H}^{\mathrm{BFGS}} \\ \hline \mathbf{H}^{\mathrm{BFGS}} & \mathbf{H}^{\mathrm{BFGS}} \end{array}\right) \in \mathbb{R}^{(n_{\mathbf{p}}+n_{\mathbf{d}}) \times (n_{\mathbf{p}}+n_{\mathbf{d}})}, \tag{15}$$

where we insert the analytical Hessian $\nabla_{\overline{\mathbf{p}}}^2 Y(\overline{\mathbf{p}}, \mathbf{d}) \in \mathbb{R}^{n_{\mathbf{p}} \times n_{\mathbf{p}}}$ from (8) into the BFGS formulation (14). The mixed strategy can also be necessary, if the gradient or Hessian of the pdf cannot be computed in closed form.

4 Numerical Results

As benchmark problem we consider a simple dielectrical waveguide with two uncertain geometrical parameters p_1 (length of the inlay) and p_2 (length of the

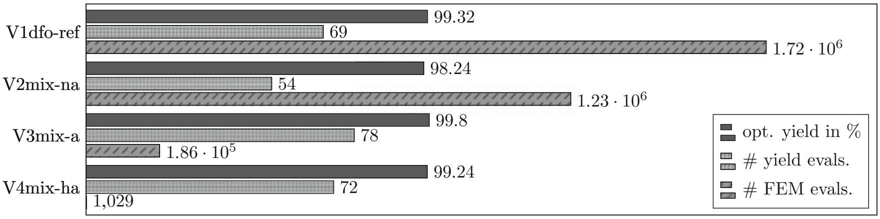


Fig. 1 Comparison of different methods for yield optimization

offset) and two deterministic material parameters d_1 and d_2 with impact on the relative permittivity and permeability. The uncertain parameters are assumed to be independent truncated Gaussian distributed with truncation at ±3 mm in order to avoid unphysical values. Thus, the parameters and their initial values for optimization are given by

$$\overline{\mathbf{p}}^0 = [9, 5] \,, \; \boldsymbol{\Sigma} = \operatorname{diag}\left(\left[0.9^2, 0.9^2\right]\right) \text{ and } \mathbf{d}^0 = [1, 1] \,. \tag{16}$$

The range parameter is the angular frequency. The QoI is the scattering parameter (S-parameter), i.e., for its calculation the electric field formulation of Maxwell has to be solved numerically with FEM. We consider the pfs

$$Q_r(\mathbf{p}) \leq -24\,\text{dB} \quad \forall r \in T_r = [2\pi 6.5, 2\pi 7.5] \;\text{ in GHz.} \tag{17}$$

The frequency range T_r is discretized into 11 equidistant frequency points. For each of these points, the inequality in (17) has to be fulfilled. For more details regarding this example we refer to [9] and [3]. In the optimization we set $\sigma_{\mathrm{MC}}^{\max} = 0.01$, which implies $N_{\mathrm{MC}} = 2500$ in the non-adaptive method. In the adaptive Newton-MC we set $N_{\mathrm{MC}}^0 = 100$ and increase it if necessary. The initial yield value is $Y_{\mathrm{MC}}^0 = 42.8\%$. We compare four methods to maximize the yield of this waveguide:

- V1dfo-ref: reference solution – problem solved with classic MC for estimation and the derivative free optimization (DFO) solver Py-BOBYQA [2]
- V2mix-na: mixed strategy proposed in Sect. 3 with classic MC for estimation and non-adaptive Newton method for optimization
- V3mix-a: mixed strategy proposed in Sect. 3 with classic MC for estimation and adaptive Newton-MC for optimization
- V4mix-ha: mixed strategy proposed in Sect. 3 with Hybrid-GPR approach [4] for estimation and adaptive Newton-MC for optimization

We consider three aspects of these methods: the optimal yield they achieve, the number of objective function (i.e. yield) calls they require and the number of FEM evaluations (to solve the QoI). The results are shown in Fig. 1.

All methods achieve an improvement of the yield by more than 55% to values between 98.4% (V2mix-na) and 99.8% (V3mix-a), with the optimal solutions

$$\begin{aligned}
&\text{V1dfo-ref:} && \overline{\mathbf{p}}^{\mathrm{opt}} = [10.94, 5.22] \text{ and } \mathbf{d}^{\mathrm{opt}} = [0.44, 1.19]\\
&\text{V2mix-na:} && \overline{\mathbf{p}}^{\mathrm{opt}} = [10.56, 4.52] \text{ and } \mathbf{d}^{\mathrm{opt}} = [0.39, 1.16]\\
&\text{V3mix-a:} && \overline{\mathbf{p}}^{\mathrm{opt}} = [\ \ 9.87, 4.92] \text{ and } \mathbf{d}^{\mathrm{opt}} = [0.2, 0.125]\\
&\text{V4mix-ha:} && \overline{\mathbf{p}}^{\mathrm{opt}} = [10.86, 5.22] \text{ and } \mathbf{d}^{\mathrm{opt}} = [0.44, 1.09]
\end{aligned}$$

In the non-adaptive case (V1dfo-ref and V2mix-na) the number of yield evaluations correlates strongly with the number of FEM evaluations. The mixed strategy from Sect. 3 (V2mix-na) needs more than 20% less yield and FEM evaluations than the reference DFO solver Py-BOBYQA (V1dfo-ref). When introducing the adaptive Newton-MC (V3mix-a and V4mix-ha), the number of yield evaluations increases, which can be explained by less accurate descent directions due to noisier yield estimations because of smaller MC sample sets. Nevertheless, the total computational effort, i.e., the number of FEM evaluations, decreases, since the yield evaluations are run with smaller MC sample sets and are thus less expensive. By not applying classic MC for yield estimation, but a hybrid approach based on GPR surrogates (V4mix-ha), the computational effort again can be reduced by a factor of 180 compared to classic MC (V3mix-a), by 1191 compared to the non-adaptive strategy (V2mix-na) and by 1671 compared to the DFO reference (V1dfo-ref).

5 Conclusion

We proposed a new mixed approach to solve yield optimization problems with deterministic and uncertain optimization variables. Only for the uncertain parameters, analytical gradient and Hessian information is available. Thus, a mixed strategy with analytical and numerical (finite differences and BFGS updates) derivatives has been used. Numerical results show better efficiency than a common derivative free optimization solver. Future research will deal with implementing the adaptive strategy and available gradient information into an originally derivative free solver.

Acknowledgments The work of Mona Fuhrländer is supported by the Graduate School CE within the Centre for Computational Engineering at Technische Universität Darmstadt.

References

1. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**(3), 1005–1034 (2007).
2. Cartis, C., Fiala, J., Marteau, B., Roberts, L.: Improving the flexibility and robustness of model-based derivative-free optimization solvers. ACM Trans. Math. Soft. **45**(3), 1–41 (2019)
3. Fuhrländer, M., Georg, N., Römer, U., Schöps, S.: Yield optimization based on adaptive Newton-Monte Carlo and polynomial surrogates. Int. J. Uncert. Quant. **10**(4), 351–373 (2020).

4. Fuhrländer, M., Schöps, S.: A blackbox yield estimation workflow with Gaussian process regression applied to the design of electromagnetic devices. J. Math. Ind. **10**(25), 1–17 (2020).
5. Gallimard, L.: Adaptive reduced basis strategy for rare-event simulations. Int. J. Numer. Meth. Eng. (1), 1–20 (2019).
6. Graeb, H.E.: Analog Design Centering and Sizing. Springer, Dordrecht (2007)
7. Hammersley, J.M., Handscomb, D.C.: Monte Carlo methods. Methuen & Co Ltd, London (1964)
8. Li, J., Xiu, D.: Evaluation of failure probability via surrogate models. J. Comput. Phys. **229**(23), 8966–8980 (2010).
9. Loukrezis, D.: Benchmark models for uncertainty quantification (2019). https://github.com/dlouk/UQ_benchmark_models/tree/master/rectangular_waveguides/debye1.py
10. Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
11. Ulbrich, M., Ulbrich, S.: Nichtlineare Optimierung. Birkhäuser (2012)

Thermomechanical Modelling for Industrial Applications

Nirav Vasant Shah, Michele Girfoglio, and Gianluigi Rozza

Abstract In this work we briefly present a thermomechanical model that could serve as starting point for industrial applications. We address the non-linearity due to temperature dependence of material properties and heterogeneity due to presence of different materials. Finally a numerical example related to the simplified geometry of blast furnace hearth walls is shown with the aim of assessing the feasibility of the modelling framework.

1 Introduction

Thermomechanical models are widely used in many practical applications [1, 2]. We refer to the case of one-way coupling between thermal and mechanical fields, where the temperature can be computed *in advance*, it being independent of the displacement, and used *afterwards* to compute the displacement. Finite Element Method (FEM) [3] is adopted to obtain these fields by solving the weak formulation of the governing equations. FEM based thermomechanical models have been successfully used for the investigation of thermomechanical phenomena arising in blast furnace [4–6]. In this work, we are going to take a step forward with respect to what done in [5] by considering the temperature dependence of material properties (that introduces a nonlinearity in the thermal model) and presence of different materials (at which one could refer to as heterogeneous material).

N. V. Shah (✉) · M. Girfoglio · G. Rozza
Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy
e-mail: shah.nirav@sissa.it; michele.girfoglio@sissa.it; gianluigi.rozza@sissa.it

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_28

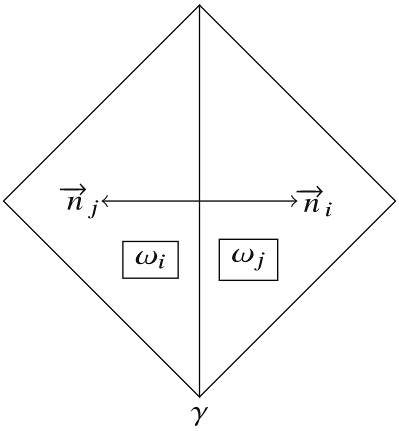
2 Mathematical Model

The blast furnace hearth is made up of several zones: ceramic cup, carbon block, steel shell. Each zone has different design requirement depending on the type of environment to which it is exposed. Ceramic cup is required to withstand high temperature due to direct contact with the molten metal. Carbon blocks are expected to reduce accumulation of excess heat. Steel shell is required to have sufficient mechanical strength to sustain the forces from other components. The reader is referred to [5, 6] for an illustration of the general layout of a blast furnace.

At the aim to consider a structure constructed using assembly of different materials, we refer to a domain ω divided into different n_{su} non-overlapping subdomains $\{\omega_i\}_{i=1}^{n_{su}}$:

$$\bar{\omega} = \bigcup_{i=1}^{n_{su}} \bar{\omega}_i \ , \ \omega_i \cap \omega_j = \emptyset \ , \ i \neq j \ . \tag{1}$$

We refer to the interface between two subdomains $\gamma = \partial\omega_i \cap \partial\omega_j \ , \ i \neq j$ as shown in Fig. 1 (left). The subdomains ω_i and ω_j are related to different materials. The temperature T and the heat flux $\overrightarrow{q} \cdot \overrightarrow{n}$, as well as the displacement $\overrightarrow{u}$ and the stress vector $\boldsymbol{\sigma}\overrightarrow{n}$ are continuous along the interface γ as reported in Fig. 1 (right). Let $k^{(i)}$, $E^{(i)}$, $\nu^{(i)}$, $\alpha^{(i)}$ respectively be the temperature dependent thermal conductivity, Young's modulus, Poisson's ratio and thermal expansion coefficient corresponding to the material of the subdomain ω_i. In the current analysis, we consider $\nu^{(i)}$ and $\alpha^{(i)}$ constant with respect to the temperature. So, for $x \in \omega_i$, we have:



$$\gamma = \partial\omega_i \cap \partial\omega_j$$
$$T_i|_\gamma = T_j|_\gamma$$
$$\overrightarrow{q}_i \cdot \overrightarrow{n}_i|_\gamma = -\overrightarrow{q}_j \cdot \overrightarrow{n}_j|_\gamma$$
$$\overrightarrow{u}_i|_\gamma = \overrightarrow{u}_j|_\gamma$$
$$\boldsymbol{\sigma}_i\overrightarrow{n}_i|_\gamma = -\boldsymbol{\sigma}_j\overrightarrow{n}_j|_\gamma$$

Fig. 1 Close-up view of two subdomains (left) and continuity conditions through the interface between the two subdomains (right)

$$k(T,x) = k^{(i)}(T)\ ,\ E(T,x) = E^{(i)}(T)\ ,\ \nu(T,x) = \nu^{(i)}\ ,\ \alpha(T,x) = \alpha^{(i)}\ .$$

We use the piecewise spline interpolation [7] to approximate thermal conductivity and Young's modulus based on their estimates (typically experimental data) related to certain discrete temperature values:

$$\text{if } T_a \le T \le T_b\ ,\ k^{(i)}(T) = a_{0,k}^{(i)}T^2 + b_{0,k}^{(i)}T + c_{0,k}^{(i)}\ ,\ E^{(i)}(T) = a_{0,E}^{(i)}T^2 + b_{0,E}^{(i)}T + c_{0,E}^{(i)}\ ,$$

$$\text{if } T_b \le T \le T_c\ ,\ k^{(i)}(T) = a_{1,k}^{(i)}T^2 + b_{1,k}^{(i)}T + c_{1,k}^{(i)}\ ,\ E^{(i)}(T) = a_{1,E}^{(i)}T^2 + b_{1,E}^{(i)}T + c_{1,E}^{(i)}\ .$$

We consider a thermomechanical problem described in the cylindrical coordinate system (r, y, θ). In many real-world applications the variation of domain geometry as well as loads and heat fluxes with respect to the angular coordinate θ could be neglected. Under such conditions, it is reasonable to apply axisymmetric hypothesis. Then in the absence of source terms the energy and momentum conservation equations in strong formulation endowed with proper boundary conditions referred to a domain ω can be stated as follows:

$$\text{Thermal model} : -\frac{1}{r}\frac{\partial}{\partial r}\left(rk\frac{\partial T}{\partial r}\right) - \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) = 0\ ,\ \text{in } \omega\ , \tag{2a}$$

$$\text{Neumann boundary} : (-k(T,x)\nabla T)\cdot\overrightarrow{n} = 0 \text{ on } \Gamma_N^T \subset \partial\omega\ , \tag{2b}$$

$$\text{Convection boundary} : (-k(T,x)\nabla T)\cdot\overrightarrow{n} = h(T - T_R) \text{ on } \Gamma_R^T \subset \partial\omega\ . \tag{2c}$$

$$\text{Mechanical model} : \frac{\partial\sigma_{rr}}{\partial r} + \frac{\partial\sigma_{ry}}{\partial y} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r} = 0\ ,\ \text{in } \omega\ , \tag{3a}$$

$$\frac{\partial\sigma_{ry}}{\partial r} + \frac{\partial\sigma_{yy}}{\partial y} + \frac{\sigma_{ry}}{r} = 0\ ,\ \text{in } \omega\ , \tag{3b}$$

$$\text{Applied force} : \boldsymbol{\sigma}\overrightarrow{n} = \overrightarrow{g}\ ,\ \text{on } \Gamma_N^u \subset \partial\omega\ , \tag{3c}$$

$$\text{Bilateral frictionless contact} : \overrightarrow{u}\cdot\overrightarrow{n} = 0\ ,\ \overrightarrow{\sigma}_t = \overrightarrow{0}\ ,\ \text{on } \Gamma_c^u \subset \partial\omega\ . \tag{3d}$$

The temperature field T and displacement field $\overrightarrow{u} = [u_r\ u_y]$ are the unknown quantities of interest. Convection coefficient h, convection temperature T_R and boundary force $\overrightarrow{g}$ are specified data. The relevant material properties include thermal conductivity k, Young's modulus E, Poisson's ratio ν and thermal expansion coefficient α. The normal vector $\overrightarrow{n}$ is considered to be pointing outwards. The shear stress $\overrightarrow{\sigma}_t$ is related to the stress tensor $\boldsymbol{\sigma}$ as:

$$\overrightarrow{\sigma}_t = \boldsymbol{\sigma}\overrightarrow{n} - \sigma_n\overrightarrow{n}\ ,\ \text{where } \sigma_n = (\boldsymbol{\sigma}\overrightarrow{n})\cdot\overrightarrow{n}\ .$$

If T_0 is the known reference temperature, axisymmetric stress-strain relationship, in vector notation, can be expressed as,

$$\{\boldsymbol{\sigma}(\overrightarrow{u})[T]\} = \boldsymbol{C}\{\boldsymbol{\varepsilon}(\overrightarrow{u})\} - \frac{E}{(1-2\nu)}\alpha(T-T_0)\{\boldsymbol{I}\}\,,$$

where $\boldsymbol{\varepsilon}$ is the strain tensor, $\boldsymbol{I}$ is the identity matrix and

$$\boldsymbol{C} = \frac{E}{(1-2\nu)(1+\nu)}\begin{pmatrix} 1-\nu & \nu & \nu & 0 \\ \nu & 1-\nu & \nu & 0 \\ \nu & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2} \end{pmatrix}.$$

We introduce weighted Sobolev spaces, $L^2_r(\omega)$ and $H^1_r(\omega)$ [8]:

$$L^2_r(\omega) = \left\{\psi : \omega \mapsto \mathbb{R}\,,\ \int_\omega \psi^2 r dr dy < \infty\right\},$$

$$H^1_r(\omega) = \left\{\psi : \omega \mapsto \mathbb{R}\,,\ \int_\omega \left(\psi^2 + \left(\frac{\partial\psi}{\partial r}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2\right) r dr dy < \infty\right\},$$

and the functional spaces for temperature and displacement:

$$\mathbb{T} = \{\psi \in L^2_r(\omega) \cap H^1_r(\omega_i)\}\,,$$

$$\mathbb{U} = \{\overrightarrow{\phi} \in [L^2_r(\omega)]^2\,, \varepsilon(\overrightarrow{\phi}) \in [L^2_r(\omega_i)]^{3\times 3}\,,\ \overrightarrow{\phi}\cdot\overrightarrow{n} = 0 \text{ on } \Gamma^u_c\}\,.$$

Then the weak formulations corresponding to equations (2) and (3) are given by:

$$\sum_{i=1}^{n_{su}} \int_{\omega_i} k\nabla T : \nabla\psi r\, dr dy + \int_{\Gamma^T_R} hT\psi r\, dr dy = \int_{\Gamma^T_R} hT_R\psi r\, dr dy\,,\ \forall\psi \in \mathbb{T}\,, \tag{4}$$

$$\sum_{i=1}^{n_{su}} \int_{\omega_i} \boldsymbol{C}\{\boldsymbol{\varepsilon}(\overrightarrow{u})\} : \{\boldsymbol{\varepsilon}(\overrightarrow{\phi})\} r\, dr dy = \sum_{i=1}^{n_{su}} \int_{\omega_i} \boldsymbol{C}(T-T_0)\alpha\{\boldsymbol{I}\} : \{\boldsymbol{\varepsilon}(\overrightarrow{\phi})\} r\, dr dy + \int_{\Gamma^u_N} \overrightarrow{\phi}\cdot\overrightarrow{g}\, r\, dr dy\,,\ \forall\overrightarrow{\phi} \in \mathbb{U}\,. \tag{5}$$

3 Numerical Example

We consider the domain ω as shown in Fig. 2. It is divided in $n_{su} = 6$ subdomains. The coordinates of their vertices are reported in Table 1.

At top boundary γ_+ and symmetry boundary $\gamma_s = \partial\omega \cap (r = 0)$, Neumann boundary (2b) and bilateral frictionless contact (3d) are applied. At bottom boundary γ_-, convection boundary (2c) and bilateral frictionless contact (3d) are applied. Inner boundary γ_{sf} and outer boundary γ_{out} are convection and applied force boundaries (Eqs. (2c), (3c)). Convection coefficient h, convection temperature T_R and applied force $\overrightarrow{g}$ are reported in Table 2.

From a physical viewpoint, Neumann boundary (2b) on γ_+ refers to the adiabatic condition. On γ_{sf}, the convection boundary (2c) refers to heat transfer with liquid iron at melting point. On γ_{out} and γ_-, the convection boundary (2c) refers to heat extraction from the structure using heat exchanger. The convection coefficients h and the convection temperatures T_R, referring to the heat exchanger operating conditions, are kept constant. Bilateral frictionless contact (3d) on γ_+ and γ_- is related to no shear force from other components and restriction on normal expansion. The restriction on normal expansion on γ_+ refers to direct contact with other sections of hearth, while the restriction on normal expansion on γ_- refers to direct contact with the ground. On the inner boundary γ_{sf}, the applied forces (Eq. (3c)) refer to hydrostatic force from molten iron. Considering that the maximum hydrostatic force is exerted when the level of molten iron is y_{max}, we take into account the worst case scenario. On boundary γ_{out}, no known force occurs.

Table 3 reports thermal conductivity and Young's modulus values used for the interpolation. It should be noted that the values reported in Table 3 are typical for the blast furnace hearth materials. The exact values of material properties depend

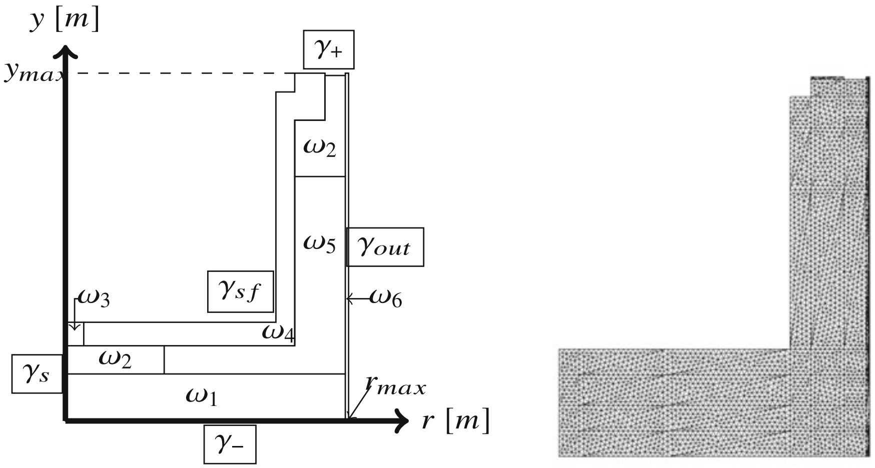


Fig. 2 Computational domain (left) and view of the mesh (right)

Table 1 Coordinates (in [m]) of the vertices of subdomains $\{\omega_i\}_{i=1}^{6}$ (see Fig. 2)

ω_1	r	0	5.9501	5.9501	2.1	0	/	/	/	/	/	/	/	/
	y	0	0	1	1	1	/	/	/	/	/	/	/	/
ω_2	r	0	2.1	2.1	0.39	0	4.875	5.9501	5.9501	5.5188	5.5188	4.875	/	/
	y	1	1	1.6	1.6	1.6	5.2	5.2	7.35	7.35	7	7	/	/
ω_3	r	0	0.39	0.39	0	/	/	/	/	/	/	/	/	/
	y	1	1	1.6	1.6	/	/	/	/	/	/	/	/	/
ω_4	r	0.39	2.1	4.875	4.875	4.875	5.5188	5.5188	5.5188	4.875	4.875	4.475	4.475	0.39
	y	1.6	1.6	1.6	5.2	6.4	6.4	7.35	7.4	7.4	7	7	2.1	2.1
ω_5	r	2.1	5.9501	5.9501	4.875	4.875	2.1	/	/	/	/	/	/	/
	y	1	1	5.2	5.2	1.6	1.6	/	/	/	/	/	/	/
ω_6	r	5.9501	5.9501	5.9501	5.9501	5.9501	6.0201	6.0201	/	/	/	/	/	/
	y	0	1	5.2	7.35	7.4	7.4	0	/	/	/	/	/	/

Table 2 Convection coefficients and temperatures, applied forces at domain boundaries

Boundary	γ_-	γ_{out}	γ_+	γ_{sf}	γ_s
Convection coefficient $h\left[\frac{\mathrm{W}}{\mathrm{m^2K}}\right]$	200	200	/	2000	/
Convection temperature T_R [K]	300	300	/	1773	/
Boundary force $\vec{g}\left[\frac{\mathrm{N}}{\mathrm{m^2}}\right]$	/	$\vec{0}$	/	$-77{,}106(y_{max} - y)\vec{n}$	/

Table 3 Temperature dependent thermal conductivity and Young's modulus values used for interpolation

	Thermal conductivity $\left[\frac{\mathrm{W}}{\mathrm{mK}}\right]$							Young's modulus [GPa]					
T [K]	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	T [K]	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
293	16.07	49.35	5.3	4.75	23.34	45.6	293	10.5	15.4	58.2	1.85	14.5	190
473	15.53	24.75	5.3	4.75	20.81	45.6	573	10.3	14.7	67.3	1.92	15.0	190
873	15.97	27.06	5.3	4.75	20.99	45.6	1073	10.4	13.8	52.9	1.83	15.3	190
1273	17.23	38.24	5.3	4.75	21.62	45.6	1273	10.3	14.4	51.6	1.85	13.3	190

on the commercial grade of the material used in the final design. Table 4[1] shows the interpolation coefficients for thermal conductivity and Young's modulus. Thermal conductivity $k^{(6)}$ and Young's modulus $E^{(6)}$ are related to thin section of steel shell where the temperature variation is not significant, so we consider them constant. On the other hand, thermal conductivities $k^{(3)}$ and $k^{(4)}$ refer to refractory blocks, which are in direct contact with high temperature molten metal and are required to sustain high temperature. They show little variation with respect to the temperature and hence, it is reasonable to assume them constant. Table 5 reports the Poisson's ratio and thermal expansion coefficient values. The reference temperature T_0 is considered as 300 K.

We use Lagrange finite element with polynomial of degree 1 for displacement and temperature. The number of degrees of freedom for temperature was 4428 and for displacement was 8856. We use Newton's method to solve the nonlinear thermal model (4) with required residual tolerance of $1e-4$. For the mechanical model (5), we use the lower-upper (LU) decomposition. All the simulations were performed by using FEniCS [9].

Numerical results are shown in Fig. 3. As can be noticed, both temperature and displacement profiles do not show strong discontinuity at the interfaces. This demonstrates that the interface conditions related to heat flux and stresses (see Fig. 1) are properly formulated and incorporated in the weak formulation.

From practical viewpoint, the temperature profile is typically used to identify areas subjected to high thermal stress. In addition, the temperature profile is also used to locate critical isotherms in the domain, such as isotherm corresponding to $1150°C$, which represents the penetration of liquid iron in the blast furnace hearth. On the other hand, the displacement profile is typically used to identify areas with

[1] The coefficients in Table 4 are rounded-off to maximum two decimal points.

Table 4 Interpolation coefficients for thermal conductivity ($T_a = 293\,\text{K}, T_b = 673\,\text{K}, T_c = 1800\,\text{K}$) and for Young's modulus ($T_a = 293\,\text{K}, T_b = 823\,\text{K}, T_c = 1800\,\text{K}$)

	$a_{0,k}^{(i)}$	$b_{0,k}^{(i)}$	$c_{0,k}^{(i)}$	$a_{1,k}^{(i)}$	$b_{1,k}^{(i)}$	$c_{1,k}^{(i)}$	$a_{0,E}^{(i)}$	$b_{0,E}^{(i)}$	$c_{0,E}^{(i)}$	$a_{1,E}^{(i)}$	$b_{1,E}^{(i)}$	$c_{1,E}^{(i)}$
ω_1	1.4E-5	−1.3E-2	1.9E1	−4.7E-6	1.1E-2	1.1E1	1.2E3	−1.6E6	1.1E10	−1.2E3	2.4E6	9.2E9
ω_2	3.9E-4	−4.3E-1	1.4E2	−1.2E-4	2.5E-1	−8.7E1	−4.5E2	−2.3E6	1.6E10	9.1E3	−1.8E7	2.3E10
ω_3	/	/	5.3	/	/	5.3	−1.05E5	1.2E8	3.1E10	6.1E4	−1.5E8	1.4E11
ω_4	/	/	4.75	/	/	4.75	−7.4E2	8.8E5	1.7E9	6.5E2	−1.4E6	2.6E9
ω_5	3.9E-5	−4.4E-2	3.3E1	−1.3E-5	2.6E-2	9.2	1.3E3	8.9E5	1.4E10	−1.9E4	3.4E7	5.6E8
ω_6	/	/	45.6	/	/	45.6	/	/	1.9E11	/	/	1.9E11

Table 5 Poisson's ratio and thermal expansion coefficient values

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
$\nu^{(i)}$	0.3	0.2	0.1	0.1	0.2	0.3
$\alpha^{(i)}\ [\mathrm{K}^{-1}]$	2.3E-6	4.6E-6	4.7E-6	4.6E-6	6E-6	1.2E-5

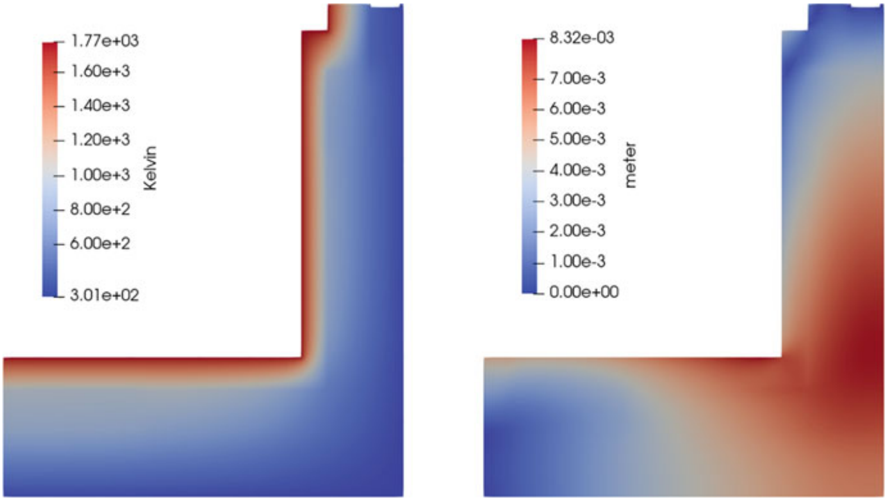


Fig. 3 Computed temperature field (left) and displacement field (right)

maximum deformation. Furthermore, the displacement field along with temperature field have direct impact on the stress field in the hearth.

4 Concluding Remarks

In this work we have addressed the development of a thermomechanical model able to describe phenomena associated to the temperature dependence of material properties (non linearity) and to the presence of different materials (heterogeneity). We expect this preliminary work could serve as starting point for thermomechanical analysis of practical problems.

Acknowledgments We acknowledge the financial support of the European Union under the Marie Sklodowska-Curie Grant Agreement No. 765374 and the partial support by the European Union Funding for Research and Innovation—Horizon 2020 Program—in the framework of European Research Council Executive Agency: Consolidator Grant H2020 ERC CoG 2015 AROMA-CFD project 681447 "Advanced Reduced Order Methods with Applications in Computational Fluid Dynamics". This work has focused exclusively on civil applications. It is not to be used for any illegal, deceptive, misleading or unethical purpose or in any military applications causing death, personal injury or severe physical or environmental damage.

References

1. Benner P., Herzog R., Lang N., Riedel I., Saak J., Comparison of model order reduction methods for optimal sensor placement for thermo-elastic models. Engrg. Optim. 51(3) (2019), 465–483.
2. Dialami N., Chiumenti M., Cervera M., and Agelet de Saracibar C., Challenges in Thermo-mechanical Analysis of Friction Stir Welding Processes. Arch. Comput. Meth. Engrg. 24 (2017), 189–225.
3. Brenner S., Scott R., (2008) The Mathematical Theory of Finite Element Methods. In: Springer-Verlag New York, 3rd ed.
4. Vázquez-Fernández S., García-Lengomín Pieiga A., Lausín-Gónzalez C., Quintela P., Mathematical modelling and numerical simulation of the heat transfer in a trough of a blast furnace. Int. J. Thermal Sci. 137 (2019), 365–374.
5. Shah N.V., Girfoglio M., Quintela P., Rozza G., García-Lengomín Pieiga A., Ballarin F., Barral P., Finite element based model order reduction for parametrized one-way coupled steady state linear thermomechanical problems. Finite Elem. Anal. Des (accepted for publication). www.arxiv.org/abs/2111.08534.
6. Shah N. V., Girfoglio M., Barral P., Rozza G., Quintela P., Lengomin A, (2022) Coupled parameterized reduced order modelling of thermomechanical phenomena arising in blast furnaces. Ph.D. thesis. Scuola Internazionale Superiore di Studi Avanzati. hdl.handle.net/20.500.11767/127929.
7. Quarteroni A., Sacco R., Saleri F., Numerical Mathematics. In: Springer-Verlag Berlin Heidelberg, 2nd ed, 2007.
8. Li H., Finite element analysis for the axisymmetric Laplace operator on polygonal domains. J. Comput. Appl. Math. 235(17) (2011), 5155–5176.
9. FEniCS Project 2019.1.0, www.fenicsproject.org.

The Virtual PaintShop: Simulation of Oven Curing

Tomas Johnson, Andreas Mark, Niklas Sandgren, Simon Sandgren, Lars Erhardsson, and Fredrik Edelvik

Abstract The modeling and simulation of oven curing in automotive paintshops is very challenging including multiple scales, turbulent air flows, thin boundary layers, large temperature gradients and long curing times. A direct brute force conjugate heat transfer simulation of an oven resolving all time and length scales would be enormously time and resource consuming. It is therefore clear that mathematical modeling must be performed, including separation of scales, and a simplification of the heat transfer coupling. We present a novel approach developed in a research project together with the Swedish automotive industry, which makes it possible to accurately simulate a curing oven with close to real time performance. The simulation results are demonstrated to be in close agreement with measurements from automotive production.

1 Introduction

There is a great need to improve the product preparation process in automotive paintshops to meet future demands on fast adaption and tailored solutions for new material combinations and products. The possibility to perform systematic simulations is then essential and would contribute to sustainable production by reducing the number of prototypes that needs to be painted, and by making it

T. Johnson · A. Mark · F. Edelvik (✉)
Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden
e-mail: tomas.johnson@fcc.chalmers.se; andreas.mark@fcc.chalmers.se; fredrik.edelvik@fcc.chalmers.se

N. Sandgren
IPS IBOFlow AB, Gothenburg, Sweden
e-mail: niklas.sandgren@ipsiboflow.com

S. Sandgren · L. Erhardsson
Scania CV AB, Södertälje, Sweden
e-mail: simon.sandgren@scania.com; lars.erhardsson@scania.com

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_29

possible to optimize the processes with respect to quality, cost and environmental impact. In earlier work we have presented novel tools for simulation of the spray and sealing processes [3, 4, 6]. The modeling and simulation of the convective ovens typically used in the automotive paintshops to cure the different paint layers is also challenging including multiple scales, turbulent air flows, thin boundary layers, large temperature gradients, and long curing times. A brute force conjugate heat transfer simulation of an oven resolving all time and length scales would be enormously time and resource consuming. Therefore, mathematical modeling is needed to obtain realistic simulations times.

We present a novel approach developed in a research project together with the Swedish automotive industry, which makes it possible to accurately simulate a curing oven in almost real time. The goal is to successfully predict the time dependent object temperature to decrease the number of physical tests that need to be carried out, especially during the production preparation phase. It also allows the oven operator to investigate possible alternative settings of the oven, e.g. flow rates and temperatures. In the approach, the individual nozzles in an oven are simulated to estimate the local nozzle Nusselt number. The Nusselt number is a dimensionless number describing the strength of heat transfer. For a complete oven, the Nusselt numbers of each nozzle are combined to model the effect of the air flow on the solid object, and thereby model the heating. Furthermore, we utilize the novel geometric routines in IBOFlow that efficiently and robustly compute the intersection between a triangular volume mesh and a hexahedral Cartesian mesh [10]. This allows us to accurately describe the solid geometry on a coarse background grid and enables the efficient solution of the heating of objects inside the oven. The novel algorithm and separation of scales approach allow us to simulate on a standard workstation. This is in contrast with previous work on simulation of oven-curing [2], where a Lattice Boltzmann solver in the fluid is coupled with a finite difference solver in the solid, which requires a large cluster to run. The simulation results are demonstrated to be in close agreement with measurements from automotive production, and they can also be utilized for multicriteria optimization [9].

2 Numerical Method

The proposed numerical method is motivated by the fact that a complete time and scale resolved simulation using the Reynolds' averaged Navier-Stokes equation together with conjugated heat transfer is very computationally demanding [2]. This is especially true since our goal is to present a method where an entire curing process, up to 1 h, can be simulated over night on a standard workstation. In this section we will describe how we solve this problem by separating the scales while preserving a physics-based approach, localize the resolved simulations, and couple the localized simulations to the full oven scale.

The numerical method has been implemented in the in-house multi-physics solver IBOFlow® [5], extending earlier available software modules employed in

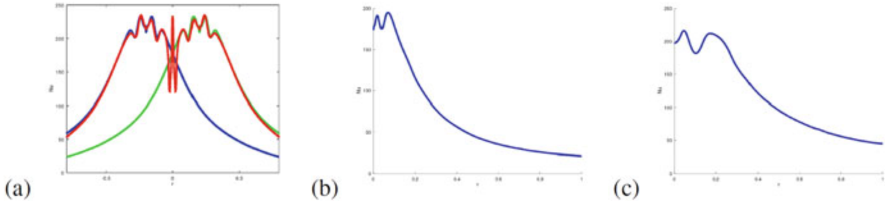


Fig. 1 Nusselt number profiles (**a**) Comparison between one nozzle and two nozzles profiles (**b**) 7 cm nozzle profile (**c**) 10 cm nozzle profile

the Virtual Paintshop. The fluid dynamics engine in IBOFlow is a co-located, segregated, incompressible Navier-Stokes solver on an octree based Cartesian mesh, which uses the SIMPLE-C method for pressure-velocity coupling. All geometries are handled with help of an immersed boundary method, for further details see [1, 5, 7, 9].

In [13] a comparison between the applicability of different turbulence models to estimate the Nusselt number for impingement heat transfer is performed. The recommendation is to use either Menter's $k - \omega$ SST or Durbin's $v^2 f$ method. We use the SST turbulence model [8], which has lower computational cost and still captures the location of the secondary peak well. The secondary peak can be seen in Fig. 1b–c and is a typical characteristic of the Nusselt number below a round jet [11–13]. The heat flux at the solid fluid interface is computed from the friction temperature and velocity. The approach is similar to the one in [11, 12].

Our approach is based on separation and localization. We localize the simulations to individual nozzles to allow us to separate the boundary layer scale from the oven scale. The scale separation contains three steps: motivation, local description, and local to global coupling.

To motivate the approach we study the interference of two nozzles. As can be seen in Fig. 1a the Nusselt number profile under two nozzles is similar to the duplication of single nozzle profiles. For the local description we generate Nusselt number profiles for a range of diameters and distances, and store all the results for varying diameters and distances in a database. Two such profiles for 7 and 10 cm nozzles are shown in Fig. 1b–c. The local to global coupling is performed by projecting the local profiles onto the object.

The body of a car or truck cab consists, to a large extent, of 1–2 mm thick sheet metal. In order to resolve such a geometry on a Cartesian grid we use the volume fraction method developed in [10], which allows us to describe the volume and area fractions locally. The method only needs a surface mesh of the object. In Fig. 2 an example of an object with two sheet metal parts is shown together with the Cartesian mesh and the volume fractions representing the object.

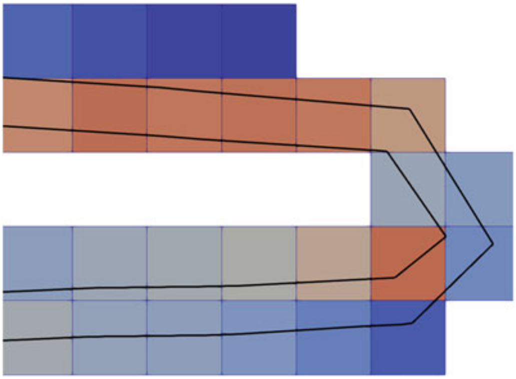


Fig. 2 Discretization of a double sheet metal part on the Cartesian mesh. The mesh is colored by the volume fraction. The conductive heat transfer is solved on the Cartesian background mesh

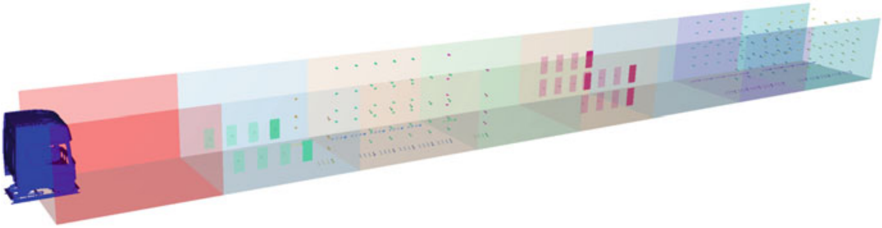


Fig. 3 The discretized oven, where the elevators are modeled as horizontal zones and the cooling zone is split into two zones due to different temperatures. The total number of discretized zones is 8 with a total of 306 circular and 24 rectangular nozzles

3 Results

To validate our approach we simulate the curing of a Scania R20H cab in a convective curing oven at the paint shop in Oskarshamn, Sweden. The oven is shown in Fig. 3. It has 306 circular and 24 rectangular nozzles. The measurements are performed on a dry cab with 7 probes attached to it. The probes are positioned to give an accurate description of the heating of the cab, including areas such as beams with thicker material. To ensure proper curing the resulting oven curves should match the specification of the paint manufacturer. In particular the minimum time above paint specific critical temperatures must be ensured.

The results of the simulation compared with the measurements are shown in Fig. 4. As seen in the figure the simulations closely capture the temperature profiles from the measurements. The point-wise mean deviation between measured and simulated temperatures for the 7 probes are shown in Table 1. The least accurate probe (2) is positioned on a thin part in the front of the cab, where position is important, and the projected fluxes give a larger error compared to a full simulation.

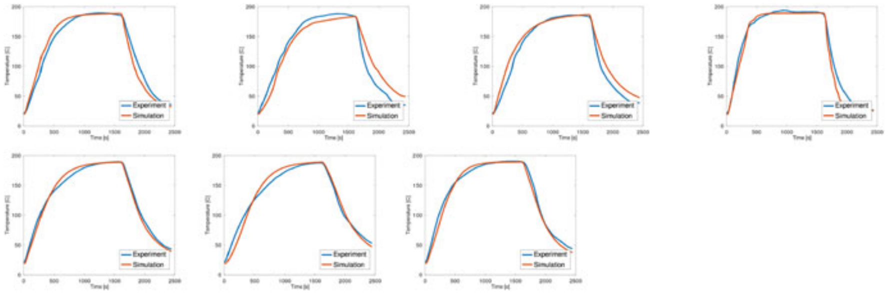


Fig. 4 Validation of the heating of the cab in the oven for 7 probes. They are shown in order from Probe 1 to Probe 7

Table 1 The point-wise mean deviation between measured and simulation temperatures for the 7 probes

Probe	1	2	3	4	5	6	7
Deviation [%]	4.2	7.2	5.0	3.5	2.5	3.6	2.8

4 Conclusions

In this paper a novel framework for simulation of convective curing ovens is presented. A validation is performed for a truck cab cured in an oven at the Scania plant in Oskarshamn. Overall the agreement between simulations and measurements is very good, almost within the measurement uncertainty. The conclusion from this and other performed case studies is therefore that the simulations can be used to predict the outcome of the process, optimize process parameters and detect areas with insufficient curing. The framework is integrated in the IPS software (www.industrialpathsolutions.com) as an oven simulation module, complementing the other virtual paintshop tools. The very efficient implementation gives a major improvement of computational speed compared to earlier approaches and makes it possible to perform detailed simulations in close to real time on a standard computer. To simulate an IR oven, that are commonly used in repair shops, would be a simple extension of the work presented here.

The standard tests carried out by the automotive manufacturers are on dry objects. This is consistent with the recommendations from the paint manufacturers. Our initial goal has therefore been to validate such tests as demonstrated in this paper. The natural next step is to include transient tracking of the paint layer thickness and solvent concentration, to allow simulations of the curing itself. Future work also includes to analyze the effect of the oven curing on the adhesive joints from hemming processes.

Acknowledgments This work was supported in part by the Swedish Governmental Agency for Innovation Systems, VINNOVA, through the FFI Sustainable Production Technology program and the projects “Virtual PaintShop - Simulation of Oven Curing” and “Virtual Verification of the Hemming Process", and in part by the Production Area of Advance at Chalmers University of Technology. The computations were partly enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

1. Andersson, T., Nowak, D., Johnson, T., Mark, A., Edelvik, F., Küfer, K.-H.: Multiobjective Optimization of a Heat-Sink Design Using the Sandwiching Algorithm and an Immersed Boundary Conjugate Heat Transfer Solver, ASME J. Heat Transfer, **140**, 102002 (2018)
2. Bhardwaj, S., Euser, R., Stadik, A., Monaco, E. et al.: High Accurate Heat Transfer Tasks on Example of Body in White Drying Process in Paint Shop, SAE Technical Paper 2019-01-0185 (2019)
3. Edelvik, F., Mark, A., Karlsson, N., Johnson, T., Carlson, J.S.: Math-Based Algorithms and Software for Virtual Product Realization Implemented in Automotive Paint Shops. In: Hömberg, D., Landry, Ch., Ghezzi, L., (eds.) Math for the Digital Factory, pp. 231–251, Springer, Berlin (2017)
4. Johnson, T., Jakobsson, S., Wettervik, B., Andersson, B., Mark, A., Edelvik, F.: A finite volume method for electrostatic three species negative corona discharge simulations with application to externally charged powder bells. J. Elstat. **74**, 27–36 (2015)
5. Mark, A., van Wachem, B. G. M.: Derivation and validation of a novel implicit second-order accurate immersed boundary method. J. Comput. Phys. **227**, 6660–6680 (2008)
6. Mark, A., Andersson, B., Tafuri, S., Engström, K., Söröd, H., Edelvik, F., Carlson, J.S.: Simulation of Electrostatic Rotary Bell Spray Painting in Automotive Paint Shops. Atomiz. Sprays **23**, 25–45 (2013)
7. Mark, A., Svenning, E., Edelvik, F.: An Immersed Boundary Method for Simulation of Flow with Heat Transfer, Int. J. Heat Mass Transf., **56**, 424–435 (2013)
8. Menter, F.R.: Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications, AIAA Journal **32**(8):1598–1605 (1994)
9. Nowak, D., Johnson, T., Mark, A., Ireholm, C., Pezzotti, F., Erhardsson, L., Ståhlberg, D., Edelvik, F., Küfer, K.-H.: Multicriteria Optimization of an Oven with a Novel ε-Constraint-Based Sandwiching Method, ASME J. Heat Transfer, **143**, 012101 (2021)
10. Svelander, F., Kettil, G., Johnson, T., Mark, A., Logg, A., Edelvik, F.,: Robust Intersection of Structured Hexahedral Meshes and Degenerate Triangle Meshes With Volume Fraction Applications. Numer. Algorithms, **77**(4), 1029–1068 (2018)
11. Tejero, F., Flaszyński, P., Szwaba, R., Telega, J.: Unsteady conjugate heat transfer analysis for impinging jet cooling, Journal of Physics: Conference Series, **760**, 012034 (2016)
12. Zhu, X.W., Zhu, L., Zhao, J.Q.: An in-depth analysis of conjugate heat transfer process of impingement jet, Int. J. Heat Mass Transf. **104**, 1259–1267 (2017)
13. Zuckerman, N., Lior, N., Impingement Heat Transfer: Correlations and Numerical Modeling, J. Heat Transf. **127**(5), 544–552 (2005)

Parameter Identification and Forecast with a Biased Model

Miracle Amadi and Heikki Haario

Abstract A well known practical issue is to ascertain how well the parameters of a model can be identified so as to allow a legitimate inference. In most cases, models are biased and may not contain all the necessary features needed to fit the data well. Employing the simplest Ross model as an example, we illustrated that parameter identifiability can be a problem of three factors: model specification, noisy data and partially observed model. Kalman filtering technique was employed in order to produce an optimal estimate of the evolving state of the system based on the model and other information such as rainfall, while simultaneously estimating the model parameters using the Kalman filter likelihood. Markov Chain Monte Carlo (MCMC) was employed as a general tool to diagnose parameter identifiability. To show the performance of the methods, an illustrative example was given with malaria data from Kalangala district, Uganda. In the end, the parameters were more or less well identified although the posterior is larger than when a synthetic data was used.

1 Introduction

Determining how well the parameters of a model can be distinctly identified with the help of available data, has been a common practical issue. When model parameters are not identifiable, there is little reason to believe that estimated values are close to the actual values. Even for noiseless data, the data can be fit arbitrarily well by different combinations of parameter values for some model/data combinations, and the uncertainties in the model parameter estimations are boundless. This follows from the fact that although some parameter estimates can be obtained from a given model, these estimates could easily be local estimates or an arbitrary set of estimates that can over-fit the observation data if proper identification of the actual values of the parameters is not done.

M. Amadi (✉) · H. Haario
LUT School of Engineering Science, LUT University, Lappeenranta, Finland
e-mail: miracle.amadi@lut.fi; heikki.haario@lut.fi

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_30

Over the years, many mathematical models have been used to provide an explicit framework for understanding malaria transmission dynamics in the human populace. The basic malaria model, now known as the classical "Ross model" was developed by Sir Ronald Ross in early 1900 [5]. The Ross model has since played a key role in the development of mosquito-borne pathogen transmission studies and has had major influence on the development of strategies for malaria control. Using two differential equations for the human and mosquito, the model presents the time evolution of the fraction of individuals in infected classes (i_h, i_m):

$$\begin{aligned} di_h &= mabi_m(1 - i_h) - i_h r \\ di_m &= aci_h(1 - i_m) - \mu i_m, \end{aligned} \tag{1}$$

where i_h and i_m represents the fractions of infected humans and mosquitoes, correspondingly, m denotes the mosquito-to-human ratio, b and c denote the transmission probabilities during mosquito contact with the human, μ is the mosquito mortality rate, a is the contact rate and r represents the recovery rate for humans. Based on benchmarks described in [6], the simpler models, such as the Ross model, appear to do a better job of matching data and heuristics than the more complex models. Here, we demonstrate how well the parameters of the simple Ross model can be identified based on available data and parameter selections. The data on reported monthly malaria cases for Kalangala district for six years (2006–2011) from Uganda, and the corresponding mean monthly rainfall data were employed in this study from World Weather Online.

2 MCMC Parameter Identification

Parameter identifiability is usually diagnosed using MCMC approach. This method is based on Bayesian inference and can be used to determine the reliability of parameter estimates as well as to quantify parameter confidence. Thus, by generating distributions of parameter values consistent with the available data, this method gives reliable estimates of model parameters (and associated uncertainties) and may be used to check whether those estimates are unique. Adaptive MCMC is used in this study since we may not be able to determine a well-working proposal distribution at the outset [2]. The Adaptive MCMC is an improved version of Metropolis-algorithm that updates the proposal covariance during the MCMC run, by using information of the previously sampled points. To evaluate the fit with the data, we utilized the cost function which returns the sum of squared differences between observations and model outputs while accounting for measurement error variance. The structure of the posterior distribution shows if the observables uniquely bound the model parameters. A helpful practice for seeing how well the chain is mixing, is to make a plot of the autocorrelation functions of the parameter chain, from which one can see the degree to which samples that are k steps away correlate with each other [2]. We would expect successive points to correlate more with each other than points further apart because in MCMC, next points are dependent on the previous points. Again, the model parameters are considered to be identifiable if the parameter values that are in the best agreement with the data are bound to a small region of the parameter space.

2.1 Factors Influencing Parameter Identifiability

2.1.1 Partially Observed Model

Even for very basic models, partial observation of state variables frequently results in structural non-identifiability of model parameters [4]. The Ross model employed for this study has a compartment for infected mosquitoes population which is hardly measurable. Thus, such data is not available for this study. One approach for tailoring the model complexity to the information content of the data is to reduce the model complexity in accordance with the available data, resulting in a reduction in the ODE system's dimension [4]. A method for addressing this problem was proposed in [7], based on the practical necessity that parameters be written as functions of the known quantities of the ODE system. In this work, considering that the presence of mosquito dynamics gives an additional degree of freedom, a reduced model which has only the infectious human compartment is proposed. Therefore, the equilibrium solution of infected mosquitoes given as

$$i_m^* = \frac{i_h}{i_h + \kappa}, \qquad \text{where} \quad \kappa = \frac{\mu}{ac}, \tag{2}$$

is plugged in and parameterised as

$$\frac{di_h}{dt} = mab\frac{i_h}{i_h + \kappa}(1 - i_h) - i_h r. \tag{3}$$

The parameter κ denotes the ration of mosquito mortality and infection rate. It can be small or large depending on the size of the infected mosquito population. In our preliminary analysis, the dynamics of the original and the reduced model are the same.

2.1.2 Interdependence and Lack of Influence of Parameters

Non-identifiability can be caused by lack of influence of a parameter on the observables, as well as interdependence among the parameters. It is obvious that if a parameter has no effect on the observables, it is not possible to determine its value. On the other hand if a change in one parameter can be matched by a corresponding

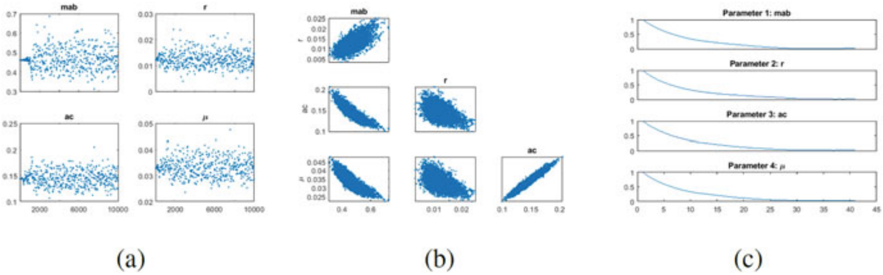


Fig. 1 MCMC results for the chosen parameterization with synthetic data: (**a**) trace plot (**b**) pairwise correlation plot (**c**) autocorrelation plot

change in another, parameter identification can be difficult, since they may not be individually identifiable [1].

Despite the fact that there is no way to absolutely establish a model's structure, unsuccessful models can be ruled out if they fail to fit the available data for any set of parameters. When the available data lack the power to constrain a model's parameters significantly, it is possible that several other models of equivalent complexity are likely to match the data well. Thus, diagnosing identifiability is a first stage in the model selection process, in which possible models are ruled out if they are unable to be bound by available data. Given that models and parameters are evaluated simultaneously, the MCMC method for detecting parameter non-identifiability may also be employed for model selection.

We tested this using synthetic data generated by adding a Gaussian noise to the output of the Ross model which was initially computed by employing values for the first set of parameters and initial conditions given in [6]. We found that all the six parameters are not well identified since the uncertainties in most of the model parameters are unbounded. However, from the nature of the mixing of the chains, it appears that some of the parameters are better expressed as products, with those products, taken as new parameters (see [1]). Following this approach, other parameterizations were evaluated, and we finally came up with one having four parameters as shown in Fig. 1. With this parameterization, the parameter chains have a very good mixing, with their associated levels of uncertainty, uniquely identified as can be seen in Fig. 1. Thus, we use this model parameterization in fitting the real data.

2.1.3 Biased Model

Bayesian identification procedure takes a long time to converge when the noise level is high [2]. Besides model parameter estimation, where the goal is to estimate static parameters, it is of interest to estimate the dynamically changing state of the system since the initial values of the system are not known. However, in some cases, the model state is not known precisely and it has to be estimated along with the parameters. State estimation in dynamical models can be done using filtering methods, where the distribution of the model state evolves along with the dynamical model and updated sequentially as new observations become available [3]. Another rationale behind filtering is that the model can have bias and may not contain all the necessary features required to fit the data well. For instance the Ross model alone does not have a provision to include the changing weather information. Thus, the filtering in this study incorporated the rainfall observations to the numerical model. The ODE was solved with the mosquito density m periodically following the rainfall with a linear model, using the time lag calculated by cross correlations and regression. The estimation of the time-lag was done by a separate analysis. Also, as the data is periodic and increasing, the Ross model has no way to fit it, but by filtering becomes yet possible.

We considered the likelihood approach of implementing parameter estimation within a data assimilation system. The likelihood of a parameter value is computed by running a state estimation procedure over a specified data set while keeping the parameter value unchanged. The likelihood is computed using the filter residuals [3]. This is similar to traditional parameter estimation, but a state estimation technique is used to "integrate out" the uncertainty in the model state. Thus, two stages are involved in order to obtain the parameter estimates:

- a filtering method for computing the posterior density for a parameter value
- a parameter estimation algorithm for obtaining the estimates.

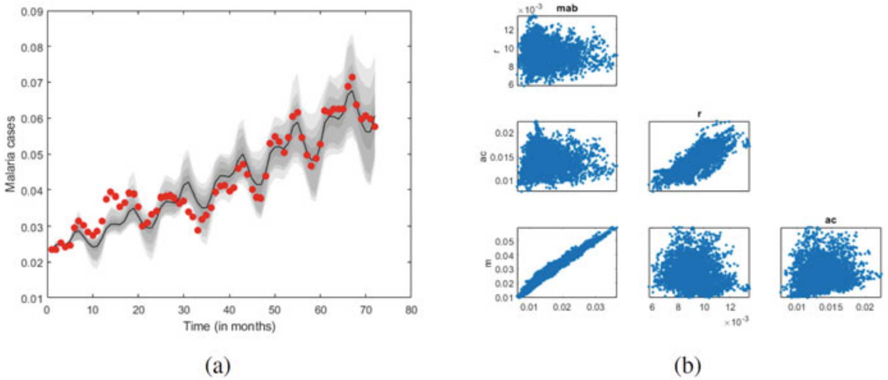
For the first task, we use the extended Kalman filter technique, since the model is non-linear. For the second task, we use the MCMC algorithm. For further reading on this approach, see [3].

3 Results

The result of the Kalman filtering done with the new ODE parameterization is given in this section. It can be seen from Fig. 2b that the parameters are properly identified. Overall, parameter identifiability improved at each step of rectifying the issues posed by the influencing factors. However, it can be seen from the plots of the two dimensional marginal distributions in Figs. 2b and 1b that the case with real data has a larger posterior density as compared to the case with synthetic data.

4 Conclusion

In general, we acknowledge that identifiability could also be a property of likelihood and suggest that the nature of the proposed model in relation to the available data be studied before embarking on full MCMC implementation. Apart from allowing for the on-line estimation of model states with relevant sources of information, the



(a)

(b)

Fig. 2 (**a**) The predictive posterior distribution of the state variable calculated from MCMC (light gray), a single prediction by MAP estimate (black bold line) and the data (red circles) (**b**) The pairwise distribution plots for the case with the real data

Kalman filtering conducted reduces uncertainties and bias, and thereby improve forecasting. The present work could be regarded as a proof of concepts that can be employed to improve parameter identifiability and forecasting.

References

1. Brouwer, A.F., Meza, R., Eisenberg, M.C.: Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. PLoS. Comput. Biol. **13**, e1005431 (2017)
2. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**, 339–354 (2006)
3. Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., Jarvinen, H.: On closure parameter estimation in chaotic systems. Nonlinear. Process. Geophys. **19**, 127-143 (2012)
4. Miao, H., Xia, X., Perelson, A.S., Wu, H.: On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM. Review **53**, 3–9 (2011)
5. Ross, R.: The prevention of malaria. London, UK (1911)
6. Wallace, D.I., Southworth, B.S., Shi, X., Chipman, J.W., Githeko, A.K.: A comparison of five malaria transmission models: benchmark tests and implications for disease control. Malar. J. **13**, 268 (2014)
7. Xia, X., Moog, C.H.: Identifiability of nonlinear systems with application to HIV/AIDS models. IEEE. Trans. Automat. Contr. **48**, 330–336 (2003)

Estimation of Time-Dependent Parameters in a Simple Compartment Model Using Covid-19 Data

Mahdi Hedayat Mahmoudi and Sara Grundel

Abstract Owing to the ongoing pandemic of COVID-19 an increased interest in epidemiological mathematical modelling arised. Several specific extensions of the classical susceptible-infected-recovered (SIR) modeling approach for the COVID-19 pandemic were developed to make forecasts. However, in all models, parameters have to be fitted on historical data. In this work we restrict ourselves to a simple model assuming however time dependent parameters. This makes sense as the parameters represent contact rate as well as recovery and death rate which are parameters that change with mutation of the virus and change in behaviour of the population. We estimate them using a Markov Chain Monte Carlo method. On the example of gender we split the model in society subgroups and estimate group specific parameters as well.

1 Introduction

The Coronavirus (COVID-19) became a major challenge during the last year [5, 11, 18], and is not yet over. The idea for this paper came from the interest in understanding the effect of different policies on the spread of the virus. Understanding this will also help in future decision-making. In particular it would be interesting to analyse the effectiveness of different methods for future pandemic handling. Different mathematical modeling approaches have been employed to simulate the disease course [14], artificial intelligence-based models [8], day-level forecasting based on time-series data [4], agent-based modeling [10], and possible others. In order to forecast rather involved models seem to be important and necessary, even though the problem of fitting the model remains tricky. Ordinary differential equation (ODE)-based models have been used for a long time to simulate the classical dynamics of epidemics [16]. In the literature various versions and

M. H. Mahmoudi · S. Grundel (✉)
Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
e-mail: mahmoudi@mpi-magdeburg.mpg.de; grundel@mpi-magdeburg.mpg.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_31

mathematical studies can be found [3, 6, 7, 13, 17]. This type of model was first proposed by Kermack and McKendrick [9] in 1927 to simulate the transmission of infectious diseases such as measles and rubella [1]. Such models assume susceptible (S), infected (I), and removed (R) fractions in a closed population and calculate the rate of changes in each fraction with ODEs [2, 19].

2 SIR Model

In this section, we use the classical SIR model, modified only to account for fatalities directly and later on a partition of each compartment by gender, which could be easily generalized to age or other society dividing features. This means the model becomes an SIRD model splitting the removed compartment into recovered (R) and dead (D) and the following ordinary differential equation and flow chart,

$$\frac{dS}{dt} = -\frac{\beta}{N} SI \quad \frac{dI}{dt} = \frac{\beta}{N} SI - (\gamma + \alpha) I \quad \frac{dR}{dt} = \gamma I \quad \frac{dD}{dt} = \alpha I \tag{1}$$

$$\begin{tikzcd}
 & & R \\
S \arrow[r, "\beta"] & I \arrow[ur, "\gamma"] \arrow[dr, "\alpha"'] & \\
 & & D
\end{tikzcd} \tag{2}$$

where β is the effective contact rate, γ the recovery rate, α the mortality rate, $N = S + I + R + D$ is the total population and t is the elapsed time from the start date. In this simple model the computation of the basic reproduction number, R_0, is the ratio of transmission and recovery plus fatal rates.

$$R_0 = \frac{\beta}{\gamma + \alpha}. \tag{3}$$

Extending this model now to different partitions of the society means that each compartment can now be split into n many separate compartments:

$$S = \bigcup_{\ell=1}^{n} S_\ell,\ I = \bigcup_{\ell=1}^{n} I_\ell,\ R = \bigcup_{\ell=1}^{n} R_\ell,\ D = \bigcup_{\ell=1}^{n} D_\ell$$

and the equations change to

$$\frac{dS_\ell}{dt} = -\sum_{k=1}^{n} \frac{\beta_{\ell k}}{N} S_\ell I_k \quad \frac{dI_\ell}{dt} = \sum_{k=1}^{n} \frac{\beta_{\ell k}}{N} S_\ell I_k - (\gamma_\ell + \alpha_\ell) I_\ell \quad \frac{dR_\ell}{dt} = \gamma_\ell I_\ell \quad \frac{dD_\ell}{dt} = \alpha_\ell I_\ell \tag{4}$$

Instead of three parameters for $n = 1$ we have in general $n^2 + 2n$ parameters, namely a size $n \times n$ matrix of parameters $\beta_{\ell k}$ and a vector γ and α. In the following we explain how a Markov Chain Monte Carlo approach can be used to estimate these parameters.

3 Markov Chain Monte Carlo Approach

The Metropolis-Hastings algorithm [12], sketched in Algorithm 1, computes a chain of parameter values $\mu_0, \ldots, \mu_N$ for given data d, a predictor $f(d, \mu)$ for the data under a parameter value μ, the size of the Markov chain N and a standard deviation σ, which represent a probability distribution of the parameter μ. This means for example that the expected value of the distribution can be computed by the mean of these values.

Algorithm 1 Metropolis-Hastings

```
Require: $N, \sigma, d$ where $d$ represents a data set
Ensure: A chain of parameter values $\mu_0, \ldots, \mu_N$
  Pick $\mu_0$
  Compute the probability of the data under this parameter by $p(d|\mu_0) \sim \exp \frac{-\|d-f(d,\mu_0)\|^2}{2\sigma}$
  Compute the probability of the parameter under the given data $p(\mu_0|d) \sim p(d|\mu_0)p(\mu_0)$ using
  a prior probability distribution of the parameter $\mu$.
  for $i < N$ do
    Pick a new parameter $\mu_j$ from a distribution $Q(\mu_j|\mu_{j-1})$
    Compute $p(\mu_j|d)$ as above
    Compute the acceptance rate $\omega(\mu_j|\mu_{j-1}) = \min(1, \frac{P(\mu_j|d)Q(\mu_j|\mu_{j-1})}{P(\mu_{j-1}|d)Q(\mu_{j-1}|\mu_j)})$
    Draw a uniform random number $U$ between 0 and 1
    if $U < \omega(\mu_j|\mu_{j-1})$ then
      $\mu_j$ is accepted and kept in the chain
    else
      $\mu_j = \mu_{j-1}$ the value from before is kept in the chain a second time.
    end if
  end for
```

4 Parameter Estimation Using RKI Data

In the following we use the Metropolis-Hastings algorithm to estimate the parameter vector $\mu = [\alpha, \beta, \gamma]$ at a given point in time. Once the particular time instance is chosen, we use the algorithm on data consisting of S, I, R, D on 20 days from that time instance on extracted from the publicly available RKI data [15]. The function f predicting the forecast of the next day based on the data of a given day and the

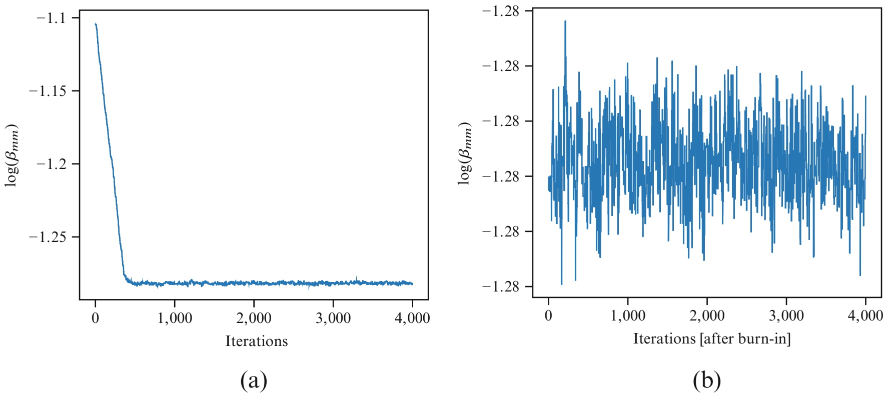


Fig. 1 Output of the Metropolis-Hastings algorithm for the sampling of β at one particular time instance. (**a**) Iterations. (**b**) Iterations [after burn-in]

parameter values is computed using an explicit Euler of (4). The prior probability distribution p is a uniform distribution and $Q(\alpha, \beta, \gamma|\alpha', \beta', \gamma')$ is a Gaussian distribution and ω is as in the algorithm but simplifies to $\omega(\beta_i^{'}|\beta_i) = min(1, \frac{P(\beta_i^{'}|d)}{P(\beta_i|d)})$ since the Gaussian is symmetric. The chain is plotted for one point in time and for the parameter β only in Fig. 1, where we can see that we have a convergence to some probability distribution in Fig. 1a and after burn-in in Fig. 1b.

Taking the mean of this chain at each point in time we plot an estimation of the time evolution of β in Fig. 2. The plot spans the time from the end of the first lockdown, where we see an increase in the effective contact rate until July 2021. We also clearly see the decrease of β as a result of the measures implemented in the fall of 2020 and the following winter. This simple analysis shows very clearly that the measures implemented had an effect on the effective contact rate.

As mentioned before we can easily extend the model to represent different groups within the society. We decided to use gender for this numerical example. This means we have for each of the four compartments S, I, R, D a split into two compartments namely into male and female and therefore a total of eight compartments and from Eq. (4) we see that we then have four values for β and two values for α and γ to a total of 8 parameters. The results of the estimation over time of these parameters with Algorithm 1 can be seen in Fig. 3 for the four values of β. The time evolution of β for male-male interaction and for female-female interaction in Fig. 3a looks very similar to the general one in Fig. 2 whereas the two βs for male-female interaction are very noisy and do not allow to extract any useful information from. In order to analyse this further a more robust approach needs to be used. This is particularly important once we use this technique on the more interesting split of the society into the different age groups.

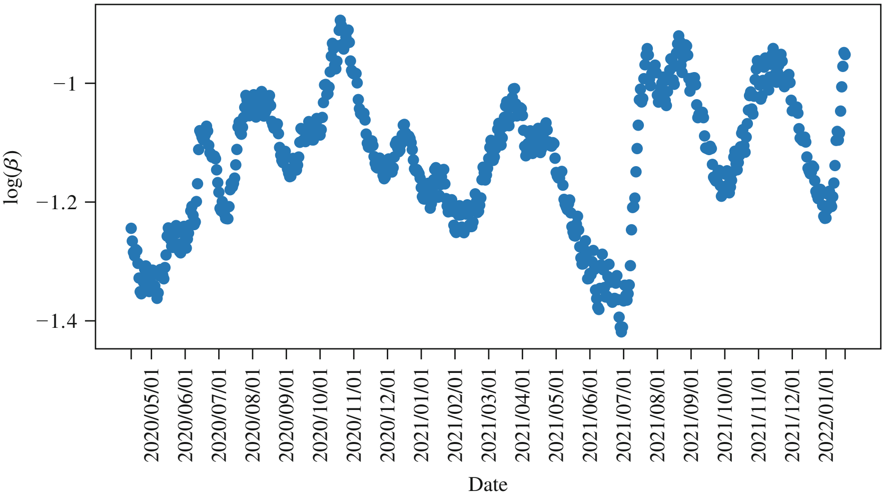
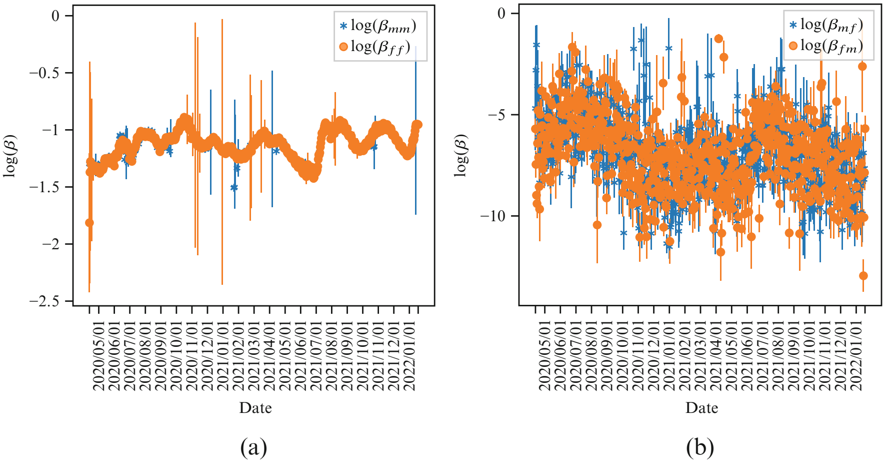


Fig. 2 Estimation of β over time as the mean of the chain computed from the Metropolis-Hastings algorithm.



(a)

(b)

Fig. 3 Estimation of the four entries of the β matrix for a gender-slit model over the time period analysed. (**a**) Male-male and female-female interaction. (**b**) Male-female and female-male interaction

5 Discussion

In the numerical example we only looked at β, and only at the mean, but with this methodology we get the full probability distribution of all parameters. It is therefore a powerful tool to estimate time dependent parameters. In future work we would want to extend this methods somewhat more to get robust results for the time evolution of the parameters and from them an estimation of the time evolution of the basic reproduction number. In a second even more interesting analysis, machine learning algorithms can then help to understand what measures influence the reproduction number to what extend.

References

1. Bacaër, N., In: A Short History of Mathematical Population Dynamics (ed. Nicolas, B.) 89–96 (Springer, London, 2011).
2. Britton, T., Stochastic epidemic models: A survey. Math. Biosci. 225, 24–35 (2010).
3. Chou, C.-S., and A. Friedman. Introduction to Mathematical Biology. Modeling, Analysis, and Simulations. Springer, 2016.
4. Elmousalami, H. H., and Hassanien, A. E., Day level forecasting for coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations. arXiv preprint arXiv:2003.07778 (2020).
5. Ferguson, N. et al., Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Imperial Coll. Lond.
6. Friedman, A., Mathematical Biology. Modeling and Analysis. CBMS Regional Conference Series in Mathematics, Vol. 127, Washington, DC, Providence, RI: American Mathematical Society, 2018.
7. Hethcote, H. W., The Mathematics of Infectious Diseases. SIAM Review 42(4) (2000), 599–653.
8. Hu, Z., Ge, Q., Jin, L., and Xiong, M. Artificial intelligence forecasting of COVID-19 in China. arXiv preprint arXiv:2002.07112 (2020).
9. Kermack W.O., McKendrick A.G., A contribution to the mathematical theory of epidemics, Proc. Royal Soc. London Ser. A Vol. 115 (1927), 700–721.
10. Kim, Y., Ryu, H., and Lee, S., Agent-based modeling for super-spreading events: A case study of MERS-CoV transmission dynamics in the Republic of Korea. Int. J. Environ. Res. Public Health 15, 2369 (2018).
11. Liu, Y., Gayle, A. A., Wilder-Smith, A., and Rocklöv, J., The reproductive number of COVID-19 is higher compared to SARS coronavirus. J. Travel Med. 27, 1–4.
12. Chib, S., and E. Greenberg. Understanding the Metropolis-Hastings algorithm. The American Statistician 49.4 (1995): 327–335.
13. Murray, J.D., Mathematical Biology. I. An Introduction, Interdisciplinary Applied Mathematics, vol 17, 3rd ed. New York: Springer-Verlag, 2002.
14. Rabajante, J. F. Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. arXiv preprint arXiv:2002.05296 (2020).
15. https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74.
16. Satsuma, J., Willox, R., Ramani, A., Grammaticos, B., and Carstea, A., Extending the SIR epidemic model. Phys. A 336, 369–375 (2004).
17. Weiss, H., A Mathematical Introduction to Population Dynamics. Rio de Janeiro: Instituto Nacional de Matemática Pura e Aplicada (IMPA), 2009.
18. Yuan, J., Li, M., Lv, G., and Lu, Z. K. Monitoring transmissibility and mortality of COVID-19 in Europe. Int. J. Infect. Dis. 95 (2020), 311–315.
19. Zhou, Y., Ma, Z., and Brauer, F. A discrete epidemic model for SARS transmission and control in China. Math. Comput. Model. 40 (2004), 1491–1506.

Comparison of Performances of Selected Forecasting Models: An Application to Dengue Data in Colombo, Sri Lanka

A. M. C. H. Attanayake, S. S. N. Perera, and U. P. Liyanage

Abstract Dengue is a one of the diseases in the world which has no exact treatment. It is rapidly spreading throughout the world by causing large number of deaths. In Sri Lanka, there is an increase of reported dengue cases over recent years. The majority of dengue cases reported in the Colombo district within the Sri Lanka. Effective dengue management strategies should be implemented to reduce the deaths from the disease. Modelling and predicting the distribution of the dengue will be useful in detecting outbreaks of the dengue and to execute controlling actions beforehand. The objective of this study is to develop an appropriate modelling technique to predict dengue cases.

To accomplish this objective, we have chosen our study area as Colombo, Sri Lanka. Seven modelling techniques, namely, Naïve, Seasonal Naïve, Random Walk with Drift, Mean Forecasting, Autoregressive Integrated Moving Average, Exponential Smoothing and TBATS were chosen in this study to model dengue data. For model development process, monthly reported dengue cases in Colombo from January 2010 to December 2018 were used and validated using the data from January to December in 2019. Mean error, root mean squared error and mean absolute percentage error measurements were used to select the most parsimonious model to predict dengue cases in Colombo. Both Exponential and TBATS models were competed in predicting dengue cases by reporting minimum error measures. Therefore, results disclosed that among the selected methods either Exponential Smoothing model or TBATS model can be used to predict dengue cases in Colombo, Sri Lanka.

A. M. C. H. Attanayake (✉) · U. P. Liyanage
Department of Statistics and Computer Science, University of Kelaniya, Kelaniya, Sri Lanka
e-mail: succ@kln.ac.lk; liyanage@kln.ac.lk

S. S. N. Perera
Research and Development Centre for Mathematical Modelling, Faculty of Science, University of Colombo, Colombo, Sri Lanka
e-mail: ssnp@maths.cmb.ac.lk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_32

1 Introduction

Dengue is one of the diseases in the world which transmits through mosquitos. When an infected dengue mosquito bites on a healthy person then the dengue virus transmits to the person. On the other way around, an uninfected mosquito bites an infected person which has dengue virus then virus may transmit to the mosquito by opening a platform to spread the disease for many people. Four serotypes were identified in the dengue virus and a person has a chance to be getting infected with all of the serotypes at different time periods. The various virus transmission ways and existence of many serotypes increase the spread of the disease through larger community.

The World Health Organization disclosed that more than 390 million people in the world infected with this complicated dengue virus annually [1]. The first dengue case reported in Sri Lanka during 1960. The majority of dengue cases reported normally in the Colombo district which is in the Western province of Sri Lanka. 10,625 of dengue cases reported in the Colombo district during the year 2019 whereas the second largest was reported in the Gampaha district which was 8432. In 2017, Sri Lanka has experienced the maximum number of dengue cases which was 186,101 throughout the country [2]. Number of deaths due to the disease and cost associated with dengue management and control increase year by year forming necessity of implementing effective and immediate actions in controlling the dengue disease.

Modelling and predicting the dengue disease play a vital role in dengue management and control by providing directions to implement right actions at the right time. Lot of researches can be found in the literature [3, 4] which were fitted to accomplished the aim of forecasting the dengue epidemic. Some of the researches [5, 6] used statistical approaches such as regression procedures and time series analysis whereas some applications used machine learning approaches [7] such as neural networks and mathematical modelling approaches [8] such as SIR (Susceptible, Infected and Recovered) and related extended models. Comparison of multiple time series modelling techniques are limited in the literature specially under the context of dengue disease in Sri Lanka. In this study, monthly dengue cases were modelled and predicted using seven selected modelling techniques; Naïve, Seasonal Naïve, Random Walk with Drift, Mean Forecasting, Autoregressive Integrated Moving Average, Exponential Smoothing and TBATS (Trigonometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal components) for the Colombo district. These techniques are some of the fundamental and most widely used techniques available in the area of time series analysis. Mean absolute error, mean absolute percentage error and root mean squared error were used to find the most parsimonious model among the selected models for predicting the dengue cases in Colombo. Availability of an accurate predicting model will lead in proposing controlling actions towards managing the disease.

2 Materials and Methods

2.1 Data Source

Monthly reported dengue cases in the district of Colombo were collected from the official web page of the Epidemiology Unit of Ministry of Health, Sri Lanka from the period of January 2010 to December 2019. Data from 2010 to 2018 were served for the development of seven models and rest of the data to check the adequacy of fit.

2.2 Forecasting Models

Following forecasting models were applied in the study:

2.3 Naïve Method

In the Naïve method, all forecasts are equal to the last observed value. This type of forecasting method will be useful if data represents a white noise. The Naïve method introduces baseline for advanced models. This method may not suitable in long term predictions.

2.4 Seasonal Naïve Method

An extended model of Naïve is called Seasonal Naïve method. If the series exhibits seasonal pattern, then this forecasting method would be appropriate. Prediction is equal to the last observed value of the same season. For an example, predictions for all future months of January are equal to the value of the last January [9].

2.5 Random Walk with Drift Method

This method is equal to the draw a line between the first and the last observations of the series and extrapolating it into the future. The name Drift has the meaning of amount of change over time. The value of drift will add the average value in order to make forecasts.

2.6 Mean Forecasting Method

This method produces forecasts which is equal to the average of all the past data. The same forecasts will produce for all of the future predictions by averaging out all of the unusual and unexpected details. Even though these same forecasts may not accurate, investigators will be able to understand the underlying process by knowing the expected value of the series.

2.7 Autoregressive Integrated Moving Average (ARIMA) Method

ARIMA is a univariate time series modelling approach which has wider applications in almost all of the fields. If the original time series exhibits non-stationarity, then the series should be converted in to a stationary series by considering seasonal and/or non-seasonal differencing. Autocorrelation and partial autocorrelation functions use to identify possible autoregressive and moving average parameters [10]. In model diagnostic checking, residuals should follow a white noise which is drawn from a constant mean and variance. In the case of violation of assumptions another model need to be investigated otherwise the selected model can be used to make predictions. To capture seasonality, SARIMA (seasonal ARIMA) models can be used.

2.8 Exponential Smoothing Method

As the name implies, this method assigns exponential weights to the observations. More recent the observation will get higher weight. There are various forms of exponential smoothings and the simplest form of the exponential smoothing is named as 'simple exponential smoothing' which applicable when the series does not contain any trend or seasonality. The double exponential smoothing method wold be appropriate if the data represent some trend. If the data represent both trend and seasonality, then Holt-winters smoothing method may be suitable. This method has two options to capture seasonality and trend as additive or multiplicative. Three smoothing parameters are in the model to capture pattern, trend and seasonality respectively. All three parameters are in between 0 and 1. By considering trend, seasonality and resulting error structures as either additive or multiplicative, varies models can be constructed and validated. 'ets' function in R software was used in finding the weights of the model.

2.9 TBATS (Trigonometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal components) Method

TBATS method is an appropriate forecasting technique if the series has complex and multiple seasonal patterns. TBATS stands for Trigonometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal components. The TBATS model fit as TBATS($\omega, p, q, r, m_1, k_1, \ldots, m_j, k_j$) where ω is the Box-Cox parameter and r is the damping parameter. The error is modelled as an ARMA (p, q) process and m_1 through m_j denote the seasonal periods used in the model and k_1 through k_j are the corresponding number of Fourier terms used for each seasonality [11].

3 Results and Discussion

The analysis was mainly performed using R software [12]. The time series plot of monthly reported dengue cases in Colombo, Sri Lanka from 2010 to 2018 is shown in Fig. 1. According to Fig. 1, the reported dengue cases varies in between the minimum value of 97 and the maximum value of 3620 other than the highest numbers of dengue cases reported in June and July of 2017 which were 5372 and 7471 cases.

The Naïve, Seasonal Naïve, Random Walk with Drift and Mean Forecasting methods were applied on the data from the period of 2010 to 2018 in order to find the forecasts for the year 2019. The forecasted values of each of the methods display in Fig. 2.

Mean forecast of dengue cases by the Mean method was 1137. Therefore, the average value of the dengue cases throughout the period of 2010–2018 is 1137 dengue cases. Naïve forecast of dengue cases was 1333. Forecasted values generated for January to December in 2019 by the Drift method were 1340, 1347, 1354, 1361, 1368, 1375, 1382, 1389, 1396, 1403, 1410 and 1417. According to the drift method forecasts increase very slowly during the year 2019. Seasonal Naïve method was able to capture the seasonality of the dengue series up to a certain level.

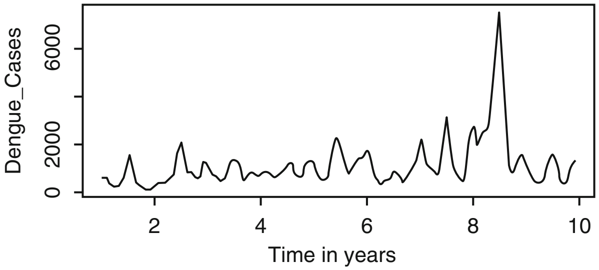


Fig. 1 Time series plot of monthly reported dengue cases in Colombo, Sri Lanka

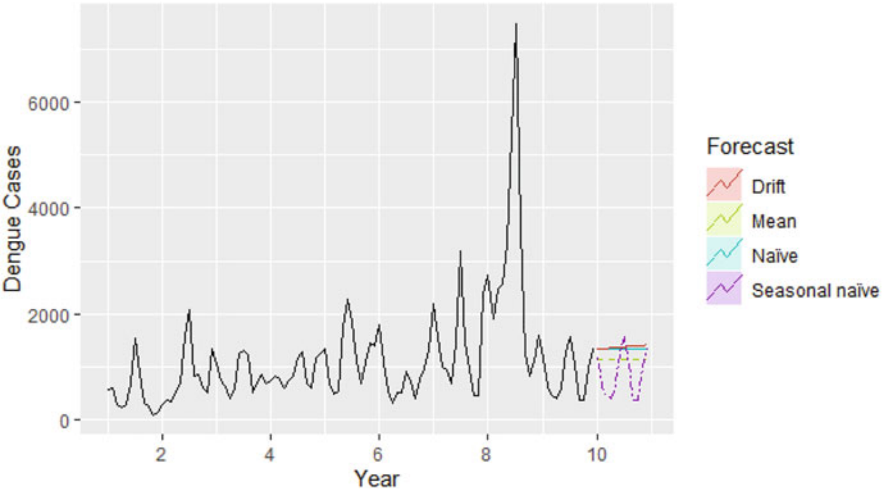


Fig. 2 Forecasts of Dengue by Naïve, Seasonal naïve, Mean and Drift Methods

Augmented Dickey Fuller (ADF) Test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test confirmed the non-stationarity of the original series at 5% significance level. Both seasonal and non-seasonal differencing overcome the non-stationarity of the series at 5% significance level. The ADF and KPSS tests confirmed the stationarity of the differenced series. By changing parameters of autoregressive and moving average components of ARIMA model, optimum model for Colombo district was found which have minimum AIC, BIC and AICc [10] measures. Then optimum model was check for the validity of assumptions. Residual analysis (error analysis) of the model represents in Fig. 3. All the assumptions of residuals satisfied by the optimum model whereas Ljung-Box test is not significant at 5%. The selected best model for Colombo is SARIMA (0,1,2) (0,1,1)12 among candidate SARIMA models. The selected best SARIMA model was used to forecast the dengue cases from January to December in 2019. Forecasted values are shown in Fig. 4 with 80% and 95% confidence intervals.

Forecasts in non-differenced scale were obtained and values match with the actual figures only in first few months of 2019 (Fig. 7).

As non-stationarity of the original dengue series was confirmed by ADF and KPSS tests both simple and double exponential smoothing techniques will not appropriate in modelling the original dengue series. Hence, Holt Winters smoothing technique was applied to model the original series of dengue cases. All possible combinations that can be considered for modelling by changing multiplicative and additive structures for all error, trend and seasonality of the series were implemented. The optimal exponential smoothing model with minimum AIC, BIC, MAE, MAPE and RMSE selected as the best smoothing model to forecast future dengue cases in Colombo. It consists with multiplicative error, multiplicative

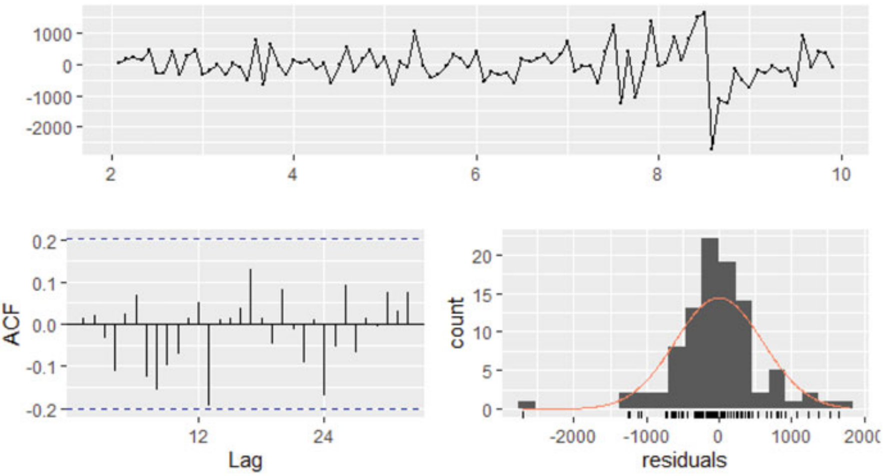


Fig. 3 Residual analysis of SARIMA (0,1,2) (0,1,1)12

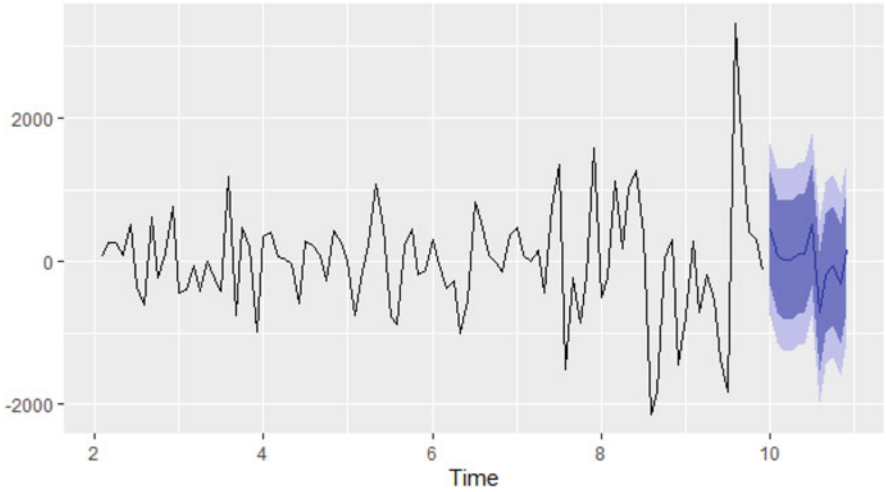


Fig. 4 Forecasts from SARIMA (0,1,2) (0,1,1)12

seasonality and additive structure for trend. Forecasted values for the year 2019 are given in Fig. 5.

The TBATS model was fit by using the 'tbats' function of the R package. The forecasts are shown in Fig. 6.

The recorded MAE, MAPE and RMSE values of each technique were summarized in Table 1. Results of the Table 1 revealed that both exponential smoothing and TBATS methods appropriate in forecasting monthly dengue cases in Colombo, Sri Lanka by reporting minimum values for MAE, MAE and RMSE error measures.

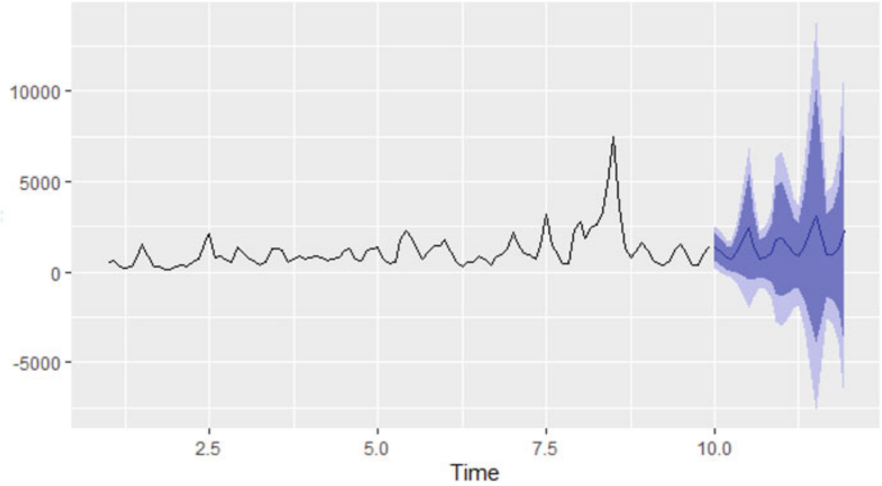


Fig. 5 Forecasts from the Best Exponential Smoothing model

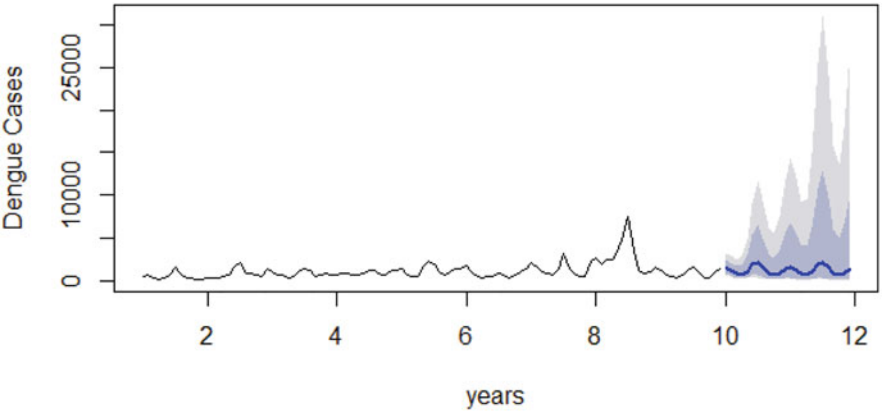


Fig. 6 Forecasts from TBATS model

Table 1 Error measures of forecasting models.

Method	MAE	MAPE	RMSE
ARIMA	411.57	231.53	600.42
Exponential smoothing	297.00	33.80	415.47
Naive	470.94	44.70	737.64
SNaive	794.84	75.87	1283.80
Drift	470.73	44.92	737.60
Mean	631.54	90.79	1005.84
TBATS	298.31	29.90	447.13

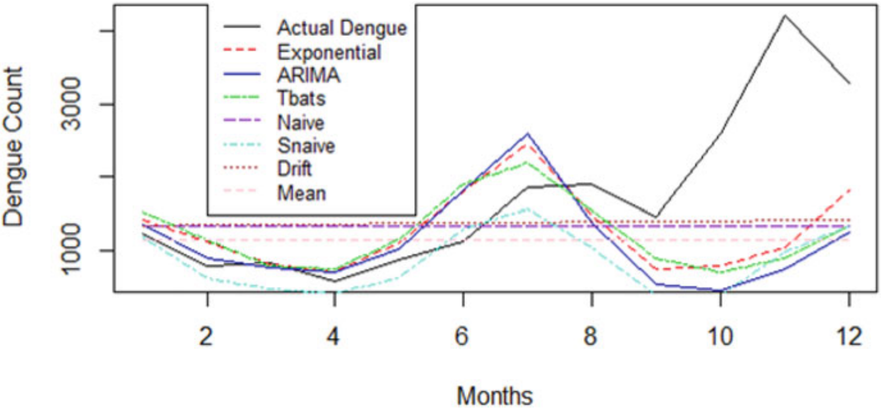


Fig. 7 Forecasts from seven methods for the year 2019

Forecasts generated by each method for the year 2019 with actual reported dengue cases in 2019 displays in Fig. 7. It can be seen by Fig. 7 that both Exponential smoothing and TBATS models were able to capture some of the patterns exists in the original series in an acceptable magnitude. Therefore, both Exponential and TBATS models were recommended for forecasting monthly dengue cases in Colombo, Sri Lanka within the other models considered in this study.

4 Conclusion

This study successfully models the monthly reported dengue cases in Colombo, Sri Lanka through seven forecasting techniques namely Naïve, Seasonal Naïve, Random Walk with Drift, Mean Forecasting, Autoregressive Integrated Moving Average, Exponential Smoothing and TBATS (Trigonometric, Box-Cox Transformation, ARMA errors, Trend and Seasonal components) with the aim of forecasting future dengue cases. Three error measures as MAE, MAPE and RMSE were used to compare the performances of the fitted seven models. The minimum error measures were reported for the Exponential smoothing and TBATS models. Therefore, results of the study disclosed that among the selected methods either Exponential Smoothing model or TBATS model can be used to predict dengue cases in Colombo, Sri Lanka. The forecasted values generated by these models may be useful in taking actions towards controlling the dengue cases in Colombo, Sri Lanka.

References

1. World Health Organization (2019), Dengue and Severe Dengue, Available via http://www.who.int/mediacentre/factsheets/fs117/en/.
2. Epidemiology Unit, Ministry of Healthcare and Nutrition, Sri Lanka (2019), Dengue Update, Available via http://www.epid.gov.lk.
3. Wickramaarachchi W.P.T.M., Perera S.S.N., Jayasinghe S., Modelling and Analysis of Dengue Disease Transmission in Urban Colombo: A Wavelets and Cross Wavelets Approach, J. Nat. Sci. Found., Sri Lanka, 43(4) (2015), 337–345.
4. Lai Y.H. (2018). The climatic factors affecting dengue fever outbreaks in southern Taiwan: an application of symbolic data analysis, Biomed. Eng. Online, 17(2), 148.
5. Attanayake A.M.C.H., Perera S.S.N., Liyanage U.P., Combining Forecasts of ARIMA and Exponential Smoothing Models, Advances and Applications in Statistics, Pushpa Publishing House, Allahabad, India, 59(2) (2019), 199–208.
6. Magda, M.M., Analysis of Multiple Linear Regression models using Symbolic Interval valued Variables, Int. J. Appl. Math. Stat. Sci. 7(2) (2018), 33–44.
7. Aburas H.M., Cetiner G., Sari M. (2010). Dengue confirmed cases prediction: A neural network model, *Expert Systems with Applications*, 37(6), 4256–4260.
8. Side S., Salmi M.N., A SIR model for Spread of Dengue Fever Disease, World Journal of Modelling and Simulation, 9(2) (2013), 96–105.
9. Wikipedia contributors. (2022). Forecasting. In Wikipedia, The Free Encyclopedia. Retrieved 10:09, January 30, 2022, Available via https://en.wikipedia.org/w/index.php?title=Forecasting&oldid=1067271864.
10. Wikipedia contributors. (2022). Autoregressive integrated moving average. In Wikipedia, The Free Encyclopedia. Retrieved 10:19, January 30, 2022, Available via https://en.wikipedia.org/w/index.php?title=Autoregressive_integrated_moving_average.
11. De Livera, A.M., Hyndman, R.J., Snyder, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing, Journal of the American Statistical Association, 106(496) (2011), 1513–1527.
12. R Core Team, (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Approaches for Going Beyond Linear Frequency Domain Powertrain Simulation

Klaus-Dieter Bauer, Josef Haslinger, Günter Offner, and Tigran Parikyan

Abstract The study of powertrain multi-body systems in time-domain can be prohibitively expensive for systems with high rotational speeds. Solving the equations of motion in frequency-domain can provide orders of magnitude faster results when omitting non-linear force components, allowing to separate the problem into independent equations for each load frequency. However, this feature is lost when accounting for non-linearities, e.g. from gear meshing. We present an iterative algorithm, that avoids coupling of frequency components by switching between frequency- and time-domain for describing the non-linear terms. Utility of the algorithm is demonstrated by studying a two-shaft model system, comparing solution by time-domain integration and by the iterative algorithm.

1 Introduction

Typical powertrain multi-body system simulations represent the mechanical system by components (e.g. rotating shafts) and joints (e.g. gear contacts) which describe the forces coupling their motion [1, 2, 9]. In automotive applications these systems often have thousands of degrees of freedom [5], such that obtaining the steady-state motion of fast-moving components by solving the equations of component motion in time-domain can take hours, as non-periodic deviations from the steady state motion (e.g. transient oscillations) may decay slowly relative to the system cycle

K.-D. Bauer (✉) · J. Haslinger
MathConsult GmbH, Linz, Austria

Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria
e-mail: kdbauer@mathconsult.co.at; josef.haslinger@mathconsult.co.at

G. Offner · T. Parikyan
AVL List GmbH, Graz, Austria
e-mail: guenter.offner@avl.com; tigran.parikyan@avl.com

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_33

rate. This problem is exacerbated for turbo chargers, which may rotate at as much as 350,000 rpm [7].

Frequency domain solution of the motion on the other hand can yield the steady-state motion in a matter of minutes or seconds for the same systems [10]. In this approach non-periodic deviations are inherently suppressed and linearization of the equations of motion decouples them into independent equation systems for each frequency of the applied external load—but only under the assumption, that forces can be represented by *time-independent* linearization coefficients.

Some joints (e.g. gear contacts) are modeled by force laws exhibiting stiffness fluctuations, which cannot be represented within this assumption. Naive extension to time-dependent stiffness coefficients would result in coupling across all frequencies and potentially the need for a denser frequency grid, increasing the computational cost by orders of magnitude again. Potentially this negates the performance gains over a time-based solution.

In this paper we discuss approaches for extending frequency-domain simulations of powertrain systems beyond the constant linear approximation while maintaining its performance advantage. In Sect. 2 we discuss the mathematical description of the problem. References [3, 4, 7] can be used for further reading. In Sect. 3 we develop the iterative solver algorithm by means of a perturbation approach. In Sect. 4 we demonstrate the application of the algorithm to a two-shaft model by means of a prototype implementation.

2 Background

The dynamics of a powertrain system modeled as rigid bodies and finite-element discretization of flexible bodies are represented by equations of motion of the form $M(z) \cdot \ddot{z} = f(t, z, \dot{z})$ [3] with a mass matrix $M(z)$ representing inertia effects, $z(t)$ the trajectory of the system, and f describing forces within and across bodies.

Given a decomposition $z(t) = z_0(t) + q(t)$ where $z_0(t)$ is an approximation of the real trajectory and $q(t)$ assumed to be small, and given a suitable choice of coordinate systems [2, 8, 9] or limiting the allowed models sufficiently, $M(z_0(t))$ becomes constant, and the force equation can be linearized into

$$M \cdot \ddot{q} + D(t) \cdot \dot{q} + K(t) \cdot q = f(t) \tag{1}$$

where $f(t) = f(t, z_0, \dot{z}_0) - M \cdot \ddot{z}_0$ contains external forces, internal stiffness and joint forces along the reference trajectory $z_0(t)$ and inertia forces for nodes described in accelerated coordinate systems. For sufficiently simple models and suitable $z_0(t)$, the matrices D, K become time independent along $z_0(t)$, resulting in a frequency domain problem[1]

[1] We assume that functions are decomposed into discrete Fourier coefficients according to $f(t) = \sum_\omega f_\omega e^{i\omega t}$.

$$\left(-\omega^2 M + i\omega D + K\right) \cdot q_\omega = f_\omega. \tag{2}$$

This equation can be solved separately for each frequency in $O(N_\omega N_q^2)$, where N_ω, N_q are size of the frequency grid and of the vector q respectively. However, in such a framework meshing effects cannot be represented. Taking into account the time-dependence $D(t)$, $K(t)$ results in an equation system, where the frequencies are coupled. The time complexity increases by a factor of N_ω.

3 Iterative Linear Solver

We thus study whether it is possible to enhance the results by iteratively applying the frequency solver and evaluating forces in time domain in between.

Applying the widely used perturbation theory approach [6], we split $K(t)$ into its time average K_0 and the time-dependent part $K_1(t)$, and introduce a perturbative expansion

$$\begin{aligned} K(t) &= K_0 + \lambda K_1(t) \qquad \text{and likewise for } D \\ q(t) &= q_0(t) + \lambda q_1(t) + \lambda^2 q_2(t) + \cdots \end{aligned} \tag{3}$$

where K_0 is a suitable constant component of $K(t)$ such as the time-average and $K_1(t)$ captures the time-dependence. Insertion into the equation of motion (1) yields a power series in λ, decomposing it into a sequence

$$\begin{aligned} &M \cdot \ddot{q}_0 + D_0 \cdot \dot{q}_0 + K_0 \cdot q_0 = f(t) \quad \text{for order } k = 0 \\ &M \cdot \ddot{q}_k + D_0 \cdot \dot{q}_k + K_0 \cdot q_k = -D_1(t) \cdot \dot{q}_{k-1} - K_1(t) \cdot q_{k-1} \quad \text{for } k \geq 1 \end{aligned} \tag{4}$$

with $f(t) = f(t, z_0, \dot{z}_0) - M \cdot \ddot{z}_0$ as before, by using, that the equation must remain valid for any strength scaling λ of the time-dependent part. The factor λ can be chosen to be 1.

For implementation it is convenient to reformulate in terms of the cumulative solution up to order n, $q^{(n)}(t) = \sum_{k=0}^{n} q_k(t)$. By summing over the iteration equation (4) up to order $k = n$, we obtain

$$M \cdot \ddot{q}^{(n)} + D_0 \cdot \dot{q}^{(n)} + K_0 \cdot q^{(n)} = f(t) - D_1(t) \cdot \dot{q}^{(n-1)} - K_1(t) \cdot q^{(n-1)} \tag{5}$$

with an initialization condition $q^{(-1)}(t) = 0$. This form allows us to introduce a damping factor $\gamma \in (0, 1]$, by setting $q^{(n)} \to (1 - \gamma)q^{(n)} + \gamma q^{(n-1)}$ after each solution step. The solution is then evaluated by:

1. Initialization
   - (a) Obtain some reference trajectory $z_0(t)$.
   - (b) Calculate $f(t)$, K_0, $K_1(t)$, D_0, $D_1(t)$ from $z_0(t)$.
   - (c) Initialize q_ω, $q(t)$, $\dot{q}(t)$ to 0.
2. Repeat for $n \geq 0$ until converged:
   - (a) Evaluate $f_{\text{rhs}}(t) = f(t) - M \cdot \ddot{z}_0 - D_1(t) \cdot \dot{q}(t) - K_1(t) \cdot q(t)$.
   - (b) Obtain $f_{\text{rhs},\omega}$ by Fourier analysis of $f_{\text{rhs}}(t)$.
   - (c) Solve $\left(-\omega^2 M + i\omega D_0 + K_0\right) q'_\omega = f_{\text{rhs},\omega}$.
   - (d) Update $q_\omega \to \gamma q_\omega + (1 - \gamma)\, q'_\omega$.
   - (e) Fourier synthesis of $q(t)$, $\dot{q}(t)$ from q_ω.

The result of step $n = 0$ corresponds to the linear frequency domain solution. The method of switching between time-domain and frequency-domain is intentionally left open. The easiest is to use Fast Fourier Transform (FFT), which requires equidistant frequency and time grids.

4 Application Example

We consider a simple model system consisting of two shafts connected by gears, where the pinion is driven by a turbine at a constant angular velocity, and an angle-dependent load $L(\alpha)$ acting on the gear shaft (see Fig. 1).

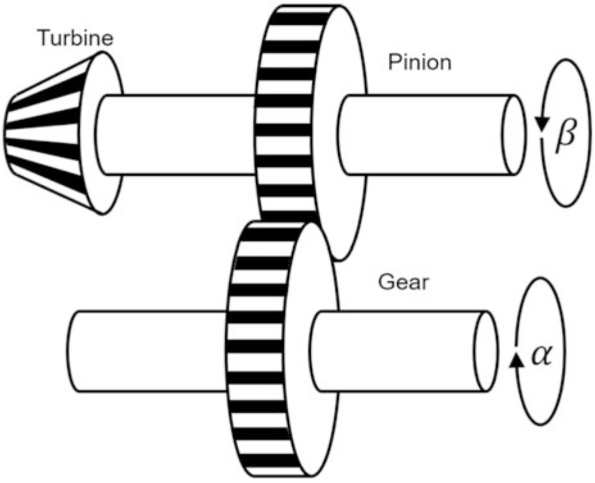


Fig. 1 Example model for demonstrating the iterative approach. The system is constrained to allow only rotations around the shaft axes and no translatory motion

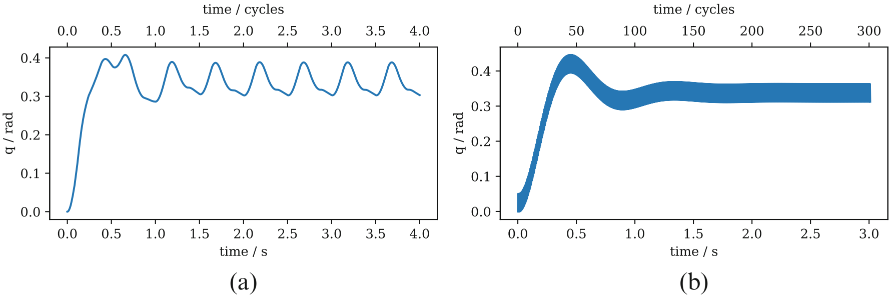


Fig. 2 Displacement $q(t)$ for the test model at (**a**) $\Omega = 60$ rpm and (**b**) $\Omega = 6000$ rpm respectively, with the load amplitude scaled as $L_1 \propto \Omega^2$ to produce similar displacement amplitudes. Deviation of the converged oscillations from a cosine-shape visible in (**a**) is caused by meshing of the gears. Since the decay time of transient terms is constant, at higher rotation speeds it takes proportionally more cycles and thus computation time to reach the steady-state behavior

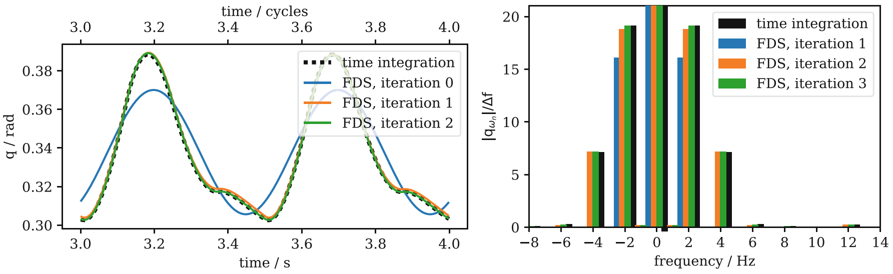
The depicted example model has a single degree of freedom α, the angular position of the gear shaft, while the trajectory of the pinion is assumed to be a uniform rotation $\beta(t) = \Omega t$. We assume a transmission ratio of 1 for simplicity, exerting a linear force $f_{\text{gear}}(\alpha) = -K(\beta)(\alpha - \beta) - D(\beta)(\dot{\alpha} - \dot{\beta})$ with $K(\beta)$, $D(\beta)$ varying periodically between single tooth and double tooth contact with 4 teeth per shaft cycle.[2] A time-dependent load of the form $f_{\text{load}}(t) = L_0 + L_1 \cos(4\Omega t)$ acts on the gear shaft. For this system, the obvious reference trajectory is $\alpha_0(t) = \Omega t$ and the displacement coordinate $q(t) = \alpha(t) - \Omega t$, resulting in an exact equation of motion

$$M\ddot{q} = f_{\text{load}}(t) - D(t)\dot{q} - K(t)q, \tag{6}$$

which we solve in time domain (Fig. 2) and by applying the algorithm described in Sect. 3 (Fig. 3), with no algorithmic damping ($\gamma = 0$).

We see that the frequency-domain algorithm reproduces the meshing effects, in this example already after one iteration beyond the linear solver, and is mostly converged with one more iteration. Repeating the simulation with 60, 600 and 6000 rpm respectively demonstrates increasingly slow convergence of the time-domain solver, with the prototype simulations taking 0.05, 0.44 and 3.94 s respectively, while the same time resolution is achieved with three iterations of the iterative frequency domain solver within a constant 0.006 s.

[2] While 4 teeth are not particularly realistic, it produces more understandable results when showing a plot over a full shaft cycle.



(a) (b)

Fig. 3 (**a**) Time domain result and (**b**) frequency spectrum obtained by a time-domain solver and the iterative frequency domain solver (FDS) algorithm from Sect. 3 for up to three iterations. The large static components ($\omega = 0$) are truncated. After one iteration ("linear solver") meshing effects are ignored entirely, visible in the spectrum (**b**) as presence of only the 2 Hz component present in the load. After only two iterations the result nearly matches the time domain solver, with only small corrections in further iterations

5 Conclusion and Outlook

We have demonstrated an iterative frequency domain solver, that provides fast solutions compared to direct time-domain integration especially at high rotation speeds, while mapping non-linear contributions to iterative solution of a linear frequency-domain problem, thus maintaining the high efficiency of a linear frequency domain solver. More studies are needed to formulate formal convergence criteria and to verify convergence for more complex models. Moreover, the algorithm should be applied to real-world application models and integrated with industrial simulation software.

Acknowledgments Part of the work of Klaus-Dieter Bauer and Josef Haslinger was supported by the Government of Upper Austria within the FTI Struktur framework.

References

1. C.B. Drab, H.W. Engl, J.R. Haslinger, G. Offner, R.U. Pfau, and W. Zulehner. Dynamic simulation of crankshaft multibody systems. Multibody System Dynamics, 22(2):133–144, 2009.
2. Th. Eizenberger. Dynamic simulation of flexible structures undergoing large gross motion, 2003.
3. M.I. Friswell, J.E.T. Penny, S.D. Garvey, and A.W. Lees. Dynamics of rotating machines. Cambridge University Press, 2010.
4. G. Genta. Dynamics of Rotating Systems. Springer Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, USA, 2005.
5. J. Haslinger, G. Offner, M. Sopouch, and B.B. Zinkiewicz. Linearized Modal Analysis of Vehicle Powertrains. In ECCOMAS Thematic Conference on Multibody Dynamics, 2017.

6. L.D. Landau and E.M. Lifshitz. Quantum Mechanics: Non-Relativistic Theory. Elsevier, October 2013.
7. H. Nguyen-Schäfer. Rotordynamics of automotive turbochargers. Springer, 2015.
8. G. Offner. Modelling of condensed flexible bodies considering non-linear inertia effects resulting from gross motions. Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics, 225(3):204–219, 2011.
9. G. Offner, Th. Eizenberger, and H.H. Priebsch. Separation of reference motions and elastic deformations in an elastic multi-body system. Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics, 220(1):63–75, 2006.
10. T. Parikyan and S. Bukovnik. Turbocharger dynamic analysis: Concept-phase simulation in frequency domain. In SIRM 2019 - 13th International Conference on Dynamics of Rotating Machined, p. 8, Copenhagen, Denmark, 2019.

Diffusion of Electron Density in Dye-Sensitized Solar Cells

Ngamta Thamwattana and Benjamin Maldon

Abstract Dye-sensitized solar cells (DSSCs) are an alternative low-cost solution to the renewable energy problem due to the use of TiO_2 as a semiconductor. Electricity generation is achieved through a series of chemical reactions designed to transport excited electrons from photosensitive dyes as a means of creating a circuit. Current modelling approach is based on the diffusion of the density of electrons in the conduction band of a DSSC's nanoporous semiconductor. In this paper, we review current models for DSSCs based on diffusion equations combining the generation and the loss of the electron density as a result of dye excitation due to sunlight and electron recombination, respectively. Further, we consider another model based on fractional diffusion equation, taking into consideration random porous network of the semiconductor TiO_2.

1 Introduction

Dye-sensitized solar cells (DSSCs) belong in the group of thin film solar cells, operating based on the photoelectrochemical processes. In general, DSSCs comprise four primary components: a photosensitive dye, a nanoporous semiconductor, an electrolyte couple and a counter electrode. Typically, a DSSC employs Ruthenium (II) photosensitive dyes, TiO_2 as nanoporous semiconductor, Iodide-Triiodide electrolyte couple and a platinum counter electrode [1, 2]. The operation starts by exposing DSSCs to sunlight, which excites dye molecules to a high energy state. This causes dye molecules to donate electrons to the nanoporous semiconductor (a process known as electron injection), which then leave the DSSC to power a load. Electrons are reintroduced through the counter electrode, which return to the photosensitive dyes through the redox electrolyte couple.

N. Thamwattana (✉) · B. Maldon
School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia
e-mail: natalie.thamwattana@newcastle.edu.au; benjamin.maldon@uon.edu.au

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_34

A dye-sensitized solar cell was proposed by O'Regan and Grätzel [1] in 1991 as an alternative cost-saving solar cells. Instead of using silicon, DSSCs adopt nanoporous titanium dioxide (TiO_2) as a semiconductor, which is much cheaper to produce. After their introduction, DSSCs have attracted much research attention including the development of new photosensitive dyes, nanoporous semiconductors, electrolyte couples and counter electrodes in order to enhance the efficiency and further lower their production costs [4].

In terms of mathematical modelling of DSSCs, many studies were based on models developed for traditional solar cell research due to similar photovoltaic principles. Another modelling approach was proposed by Södergren et al. [2] which is based on assuming diffusion of electron density in the conduction band of the nanoporous semiconductor. This assumption is also supported by Gregg [5] who stated that mathematical models for DSSCs were better informed by the influence of the photochemically induced potential over the traditional electric field approach. Since the study by Södergren et al. [2], the diffusion model for DSSCs have received further development from a simple ordinary differential equation in [2] to a fully nonlinear time-dependent partial differential equation in [6]. There are also other papers that include a system of equations to incorporate the electrolyte couple [7–9].

In this paper, we give an overview of linear and nonlinear diffusion models for electron density in the conduction band [3, 9]. This paper also considers anomalous diffusion of electron density in TiO_2 based on fractional diffusion equation and continuous-time random walk (CTRW) [10].

2 Mathematical Models and Results

In this paper, we model DSSCs based on diffusion equation. Given a DSSC of thickness d, the conduction band electron density $n(x,t)$ at position $x \in [0,d]$ and time $t \geq 0$ satisfies the diffusion equation [6] given by

$$\frac{\partial n}{\partial t} = \underbrace{D_0 \frac{\partial}{\partial x}\left[\left(\frac{n}{n_{eq}}\right)^{\beta} \frac{\partial n}{\partial x}\right]}_{diffusion} + \underbrace{\varphi \alpha e^{-\alpha x}}_{generation} \underbrace{- k_R \left(\frac{n}{n_{eq}}\right)^{\beta} \left(n - n_{eq}\right)}_{recombination}, \tag{1}$$

where D_0 is the diffusion coefficient, n_{eq} is the dark equilibrium electron density, φ is the incident photon flux, α is the absorption coefficient of the Ruthenium (II) dye, k_R is the recombination coefficient and β is the diffusion order. We note that the electron generation term is the spatially dependent, which is based on Beer-Lambert model [2, 6, 11]. The recombination term is density dependent and is referred to as a loss mechanism within solar cells, a process that hinders electricity generation [12].

Equation (1) is subject to the initial and boundary conditions:

Table 1 Numerical values of constants used in this paper [10]

Parameter	D_0	α	d	k_R	m	n_{eq}	φ
Value	10^{-11}	10^5	5×10^{-5}	4×10^{-8}	1	10^{22}	10^{21}
Unit	$\mathrm{m}^2\mathrm{s}^{-1}$	m^{-1}	m	s^{-1}	–	m^{-3}	$\mathrm{m}^{-2}\mathrm{s}^{-1}$

$$n(0,t) = n_{eq} e^{\frac{qV}{mk_BT}}, \quad \left.\frac{\partial n}{\partial x}\right|_{x=d} = 0, \quad n(x,0) = n_{eq} e^{\frac{qV}{mk_BT}}, \tag{2}$$

where q is the standard electron charge, V is the applied bias voltage of the DSSC, m is the diode ideality factor, k_B is Boltzmann's constant and T is the temperature of the DSSC. We note that for short-circuit conditions we have $V = 0$ and for open-circuit conditions, we replace the Dirichlet boundary condition at $x = 0$ with

$$\left.\frac{\partial n}{\partial x}\right|_{x=0} = 0.$$

We note that the numerical values of constants used in this paper are given in Table 1.

Next, by using scaling parameters:

$$\bar{n} = \frac{n}{n_{eq}}, \quad \bar{x} = \frac{x}{d}, \quad \bar{t} = \frac{D_0 t}{d^2},$$

the non-dimensionalised form of (1) is obtained given by

$$\frac{\partial \bar{n}}{\partial \bar{t}} = \frac{\partial}{\partial \bar{x}}\left(\bar{n}^{\beta}\frac{\partial \bar{n}}{\partial \bar{x}}\right) + \mu e^{-\nu\bar{x}} - \xi\bar{n}^{\beta}\left(\bar{n}-1\right), \tag{3}$$

where $\mu = \frac{d^2\varphi_0}{D_0 n_{eq}}$, $\nu = \alpha d$ and $\xi = \frac{k_R d^2}{D_0}$ and $\omega = \frac{qV}{mk_BT}$. Dropping the bar notation, the boundary and initial conditions for $V \neq V_{oc}$ become

$$n(x,0) = e^{\omega}, \quad n(0,t) = e^{\omega}, \quad \left.\frac{\partial n}{\partial x}\right|_{x=1} = 0,$$

and the open-circuit boundary conditions are

$$n(x,0) = e^{\omega_{oc}}, \quad \left.\frac{\partial n}{\partial x}\right|_{x=0} = 0, \quad \left.\frac{\partial n}{\partial x}\right|_{x=1} = 0,$$

where $\omega_{oc} = \frac{qV_{oc}}{mk_BT}$.

In the following two subsections, we consider special cases of (1) for linear and nonlinear diffusion equations, respectively. In Sect. 2.3, we introduce fractional diffusion model.

2.1 Linear Diffusion Model

In this subsection, we consider special case of (1) when $\beta = 0$, which corresponds to linear diffusion equation given by

$$\frac{\partial n}{\partial t} = D_0 \frac{\partial^2 n}{\partial x^2} + \varphi \alpha e^{-\alpha x} - k_R \left(n - n_{eq}\right). \tag{4}$$

Using a separation of variables approach, we obtain an analytical solution for (4) under short-circuit conditions ($V = 0$ or $\omega = 0$):

$$n(x,t) = 1 + Ae^{\sqrt{\xi}x} + Be^{-\sqrt{\xi}x} - \frac{\mu}{\nu^2 - \xi} e^{-\nu x} + \sum_{k=0}^{\infty} C_k \sin\left(\frac{(2k+1)\pi}{2} x\right) e^{-\left[\left(\frac{(2k+1)\pi}{2}\right)^2 + \xi\right]t}, \tag{5}$$

where A, B, and C_k are constants given by

$$A = -\frac{\mu\nu e^{\sqrt{\xi}-\nu} + \xi^{\frac{3}{2}}\left(e^{\omega} - 1\right) - \sqrt{\xi}\left(e^{\omega}\nu^2 - \nu^2 + \mu\right)}{\sqrt{\xi}\left(\nu^2 - \xi\right)\left(e^{2\sqrt{\xi}} + 1\right)},$$

$$B = \frac{e^{\sqrt{\xi}-\nu}\left[e^{\sqrt{\xi}+\nu+\omega}\left(\sqrt{\xi}\nu^2 - \xi^{\frac{3}{2}}\right) + \mu\nu + e^{\sqrt{\xi}+\nu}\left(\xi^{\frac{3}{2}} + \sqrt{\xi}(\mu - \nu^2)\right)\right]}{\sqrt{\xi}\left(\nu^2 - \xi\right)\left(e^{2\sqrt{\xi}} + 1\right)},$$

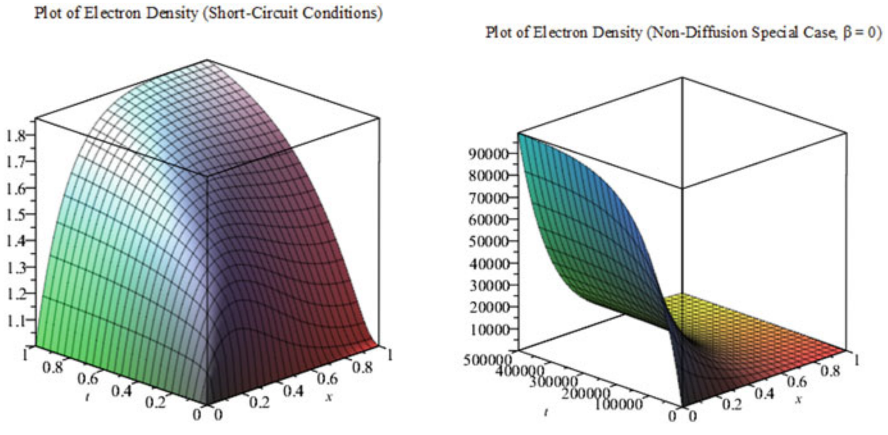
$$C_k = -2\int_0^1 \sin\left(\frac{(2k+1)\pi}{2} x\right)\left[Ae^{\sqrt{\xi}} + Be^{-\sqrt{\xi}} - \frac{\mu}{\nu^2 - \xi} e^{-\nu x}\right] dx.$$

For detailed derivation of (5) and for a solution under open-circuit conditions, we refer the readers to [9].

Further, we consider the special case when there is no diffusion term in (4). An analytical solution for this case is given by

$$n(x,t) = 1 + \frac{e^{-\nu x}}{\xi}\left(1 - e^{-\xi t}\right) - e^{-\xi t}(1 - e^{\omega}). \tag{6}$$

Plots of solutions (5) and (6) are shown in Fig. 1. With the diffusion component, the electron density rises quickly from its dark equilibrium due to the influence of the exponential source term of electron generation. The electron density continues to increase until it reaches the steady-state, as expected for photovoltaic devices. Comparison between Fig. 1a and b shows the importance of the diffusion term in the model. We note that Fig. 1 has different scale compared to Fig. 5. This is due to that Fig. 1 uses constants given in [3] while Fig. 5 adopts those presented in Table 1. We comment that the purpose of Fig. 1 is to demonstrate the significance of the diffusion term in the model.



(a) With diffusion

(b) Without diffusion

Fig. 1 (**a**) Plot of the solution (5) (with diffusion) [3] and (**b**) Plot of the solution (6) (without diffusion)

2.2 Nonlinear Diffusion Model

Here, we consider the nonlinear diffusion equation (1) when $\beta \neq 0$. In [3], both classical and nonclassical Lie symmetry methods are explored to determine analytical solutions for (1). We find solutions for special cases of no diffusion, no generation, no recombination and no generation and recombination terms. For a general case without making these assumptions, we find solutions for certain values of β when assuming certain diffusivity functions [3]. As shown in [3], for physically relevant cases, since analytical solutions are not found, we instead seek numerical solution for (1).

By adopting a forward time continuous space finite difference method (FDM) with [0, 1] as the spatial domain and 100 nodes, we plot the numerical solution $n(x, t)$ of (3) when $\beta = 1$, as shown in Fig. 2. We find that the numerical solution greatly resembles the exact solution for the linear case, suggesting that nonlinear diffusion has little effect on the profile of the solution. In Fig. 3, we compare results from our numerical scheme with those of Cao et al. [11] for $\beta = 1$. This confirms the model and the numerical scheme adopted to solve (3).

The effect of β on the solution profile is shown in Fig. 4. We can see that higher values of β lead to an overall decreased electron density. Furthermore, numerical solutions reach an equilibrium faster under increased values for β. Given that β governs the density of trap states in a DSSC (with higher values of β leading to deeper traps [6]), this result shows that the nonlinear diffusion mechanism is functioning as expected. In particular, negative values of β lead to a significantly higher electron density that has not yet reached a steady-state (unlike the corresponding nonnegative values of β). While this paper considers integer

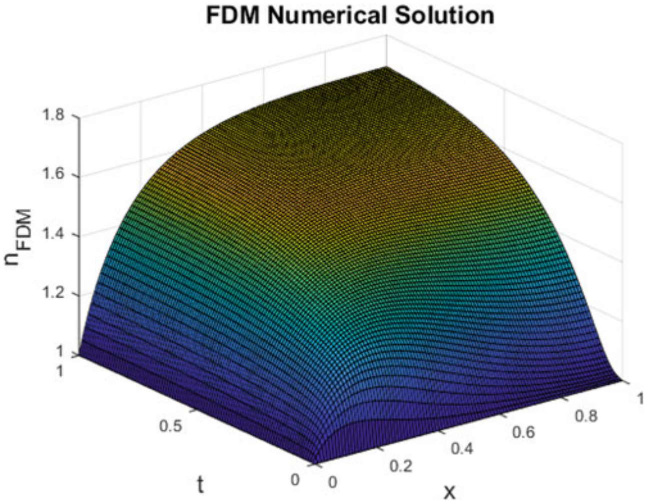
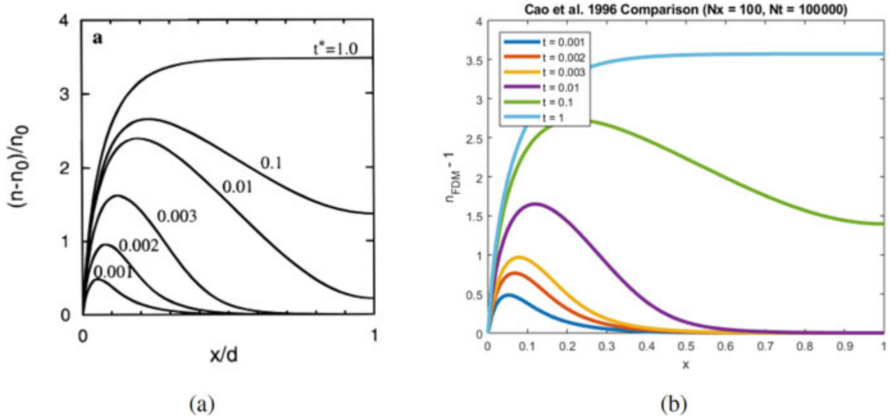


Fig. 2 Numerical solution of (3) when $\beta = 1$ [3]



(a)

(b)

Fig. 3 Comparison of numerical results between (**a**) Cao et al./ [11] and (**b**) our numerical scheme [3]

values for β between -2 and 2, the literature so far only considers $\beta = 0$ (linear case) and $\beta = 1$ [11].

2.3 Fractional Diffusion Model

Fractional diffusion model for DSSCs is proposed by Maldon and Thamwattana [10]. Their study is motivated by the connection between the fractal geometry of TiO_2 [13], which is used as a semiconductor in DSSCs, and the role of fractional

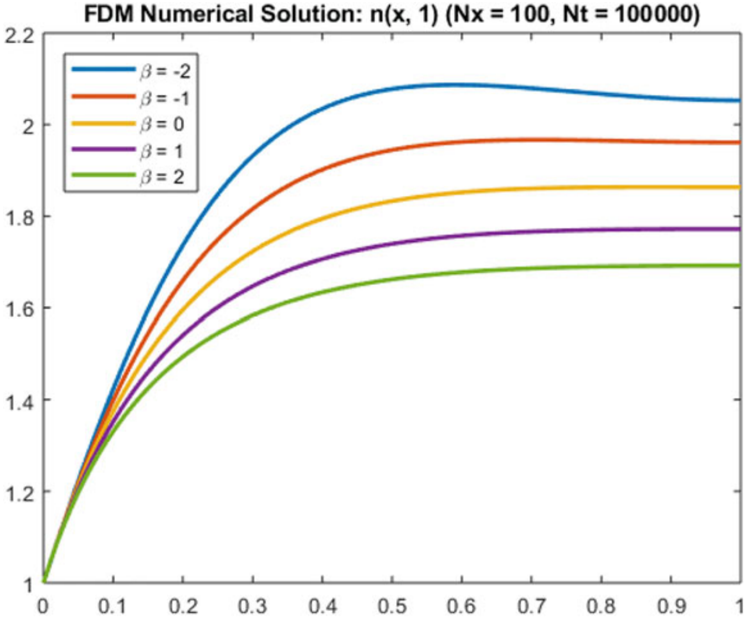


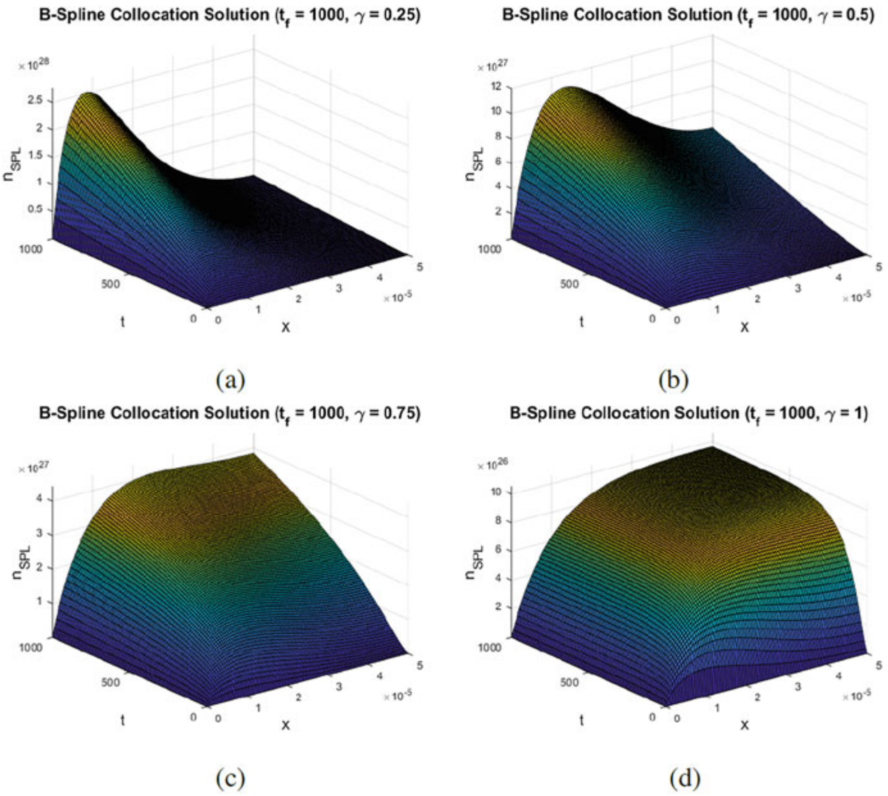
Fig. 4 Numerical solution of (3) for different values of β

derivatives in modelling diffusion in media with fractional geometry [14, 15]. Based on the general fractional reaction-diffusion equation proposed by Henry and Wearne [16], Maldon and Thamwattana [10] derive the fractional partial differential equation for DSSCs given by

$$\frac{\partial n}{\partial t} = D_0 \frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}} \frac{\partial^2 n}{\partial x^2} + \varphi \alpha e^{-\alpha x} - k_R(n - n_{eq}), \tag{7}$$

where γ is the order of fractional diffusion, noting that small γ implies slow diffusion. Other constants are as defined previously and the boundary and initial conditions are as given in (2). We note that $\gamma = 1$ leads to the standard reaction-diffusion equation.

In Fig. 5, we plot the numerical solution to (7) with final time $t_f = 1000$ for four different values for γ. Note that we use B-Spline collocation method [17] to estimate the solution over $[0, d]$ and finite difference approximation [18] to estimate the solution over time. From this figure, we see that as γ decreases the time required for the electron density to reach steady-state increases. This result is consistent with the observation that lower values for γ imply slower diffusion, based on the CTRW simulations. Further, we observe that the overall electron density is remarkably higher for the cases $\gamma = 0.5$ and 0.25 compared to $\gamma = 1$ and 0.75. This suggests that the electron density is sensitive to the order of the fractional derivative. As shown in [10], by adopting $\gamma = 0.612$ for TiO_2, the results obtained is consistent



(a) (b) (c) (d)

Fig. 5 Numerical solutions of (7) for different values of γ [10] (**a**) $\gamma = 0.25$, (**b**) $\gamma = 0.5$, (**c**) $\gamma = 0.75$ and (**d**) $\gamma = 1$.

with Benkstein et al [13]. Furthermore, Benkstein et al. [13] mentioned that higher porosity of nanoporous semiconductor (which is equivalent to slow diffusion) is not desirable as it leads to poor performance of DSSCs. Thus, the values of $\gamma > 0.5$ is more realistic when employing (7) to model diffusion of electron density in a DSSC's TiO_2 nanoporous semiconductor.

3 Conclusion

This paper gives an overview of modelling electron density in the conduction band of DSSCs based on diffusion equations. Analytical solution is presented for linear diffusion equation, which can be used to benchmark numerical calculations for nonlinear and fractional models. Using nonlinear diffusion and fractional diffusion models extend diffusion-based modelling to better quantify the performance of the nanoporous semiconductor in a DSSC. We find that the parameters β and γ have a profound effect on the electron density, signifying the important role of nanoporous TiO_2 semiconductor in the performance of DSSCs.

Acknowledgments The authors are grateful to the Australian Research Council for the funding of Discovery Project DP170102705. The authors are also grateful to Prof. Phil Broadbridge from La Trobe University, Australia, for many helpful comments and suggestions.

References

1. O'Regan, B., Grätzel, M.: A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. Nature **353**, 737–740 (1991)
2. Södergren, S., Hagfeldt, A., Olsson, J., Lindquist, S.: Theoretical models for the action spectrum and the current-voltage characteristics of microporous semiconductor films in photoelectrochemical cells. J. Phys. Chem. **98**, 5552–5556 (1994)
3. Maldon B, Thamwattana, N., Edwards, M.: Exploring nonlinear diffusion equations for modelling dye-sensitized solar cells. Entropy **22**, 248 (2020)
4. Maldon, B., Thamwattana, N.: Review of diffusion models for charge-carrier densities in dye-sensitized solar cells. J. Phys. Commun. **4**, 1–18 (2020)
5. Gregg, B.A.: Comment on "Diffusion impedance and space charge capacitance in the nanoporous dye-sensitized electrochemical solar cell" and "Electronic transport in dye-sensitized nanoporous TiO_2 solar cells - comparison of electrolyte and solid-state devices". J. Phys. Chem. B **107**, 13540 (2003)
6. Anta, J.A., Casanueva, F., Oskam, G.A.: Numerical model for charge transport and recombination in dye-sensitized solar cells. J. Phys. Chem. B **110**, 5372–53788 (2006)
7. Andrade, L., Sousa, J., Ribeiro, H.A., Mendes, A.: Phenomenological modeling of dye-sensitized solar cells under transient conditions. Sol. Energy **85**, 781–793 (2011)
8. Papageorgiou, N., Grätzel, M., Infelta, P.P.: On the relevance of mass transport in thin layer nanocrystalline photoelectrochemical solar cells. Sol. Energy Mater. Sol. Cells **44**, 405–438 (1996)
9. Maldon, B., Thamwattana, N.: An analytical solution for charge carrier densities in dye-sensitized solar cells. J. Photochem. Photobiol. A **370**, 41–50 (2019)
10. Maldon, B., Thamwattana, N.: A Fractional Diffusion Model for Dye-Sensitized Solar Cells. Molecules **25**, 2966 (2020)
11. Cao, F., Oskam, G., Meyer, G.J., Searson, P.C.: Electron transport in porous nanocrystalline TiO_2 photoelectrochemical cells. J. Phys. Chem. **100**, 17021–17027 (1996)
12. Le Bahers, T., Pauporté, T., Lainé, P.P., Labat, F., Adamo, C., Ciofini, I.: Modeling dye-sensitized solar cells: From theory to experiment. J. Phys. Chem. Lett. **4**, 1044–1050 (2013)
13. Benkstein, K.D., Kopidakis, N., van de Lagemaat, J., Frank, A.J.: Influence of the percolation network geometry on electron transport in dye-sensitized titanium dioxide solar cells. J. Phys. Chem. B **107**, 7759–7767 (2003)
14. O'Shaughnessy, B., Procaccia, I.: Analytical solutions for diffusion on fractal objects. Phys. Rev. Lett. **54**, 455–458 (1985)
15. Nigmatullin, R.: The realization of the generalised transfer equation in a medium with fractal geometry. Phys. Status Solidi B **133**, 425–430 (1986)
16. Henry, B.I., Wearne. S.L.: Fractional reaction-diffusion. Phys. A **276**, 448–455 (2000)
17. Mittal, R.C., Jain, R.K.: Cubic B-splines collocation method for solving nonlinear parabolic partial differential equations with Neumann boundary conditions. Commun. Nonlinear Sci. Numer. Simul. **17**, 4616–4625 (2012)
18. Oldham, K., Spanier, J.: The Fractional Calculus: Theory and Applications of Differentiation and Integration to Arbitrary Order. Elsevier (1974)

Modeling and Simulation of Inelastic Effects in Composite Cables

Davide Manfredo, Vanessa Dörlich, Joachim Linn, and Martin Arnold

Abstract The present work aims at describing hysteresis behaviour arising from cyclic bending experiments on cables by means of the Preisach operator. Pure bending experiments conducted in previous work show that slender structures such as electric cables behave inelastically and open hysteresis loops arise, with noticeable difference between the first load cycle and the following ones. The Preisach operator plays an important role in describing the input-output relation in hysteresis behaviours and it can be expressed as a superposition of relay operators. Here, we utilise data collected from pure bending experiments for a first approach. We introduce a mathematical formulation of the problem, and starting from the curvature of the cable specimen, we recursively define the Preisach plane for this specific case. Therefore, we derive a suitable kernel function in a way that the integration of such function over the Preisach plane results in the bending moment of the specimen.

1 Introduction

Electric cables, as those shown in Fig. 1 *left*, are complex objects due to their multi-material composition and their geometric properties. Consequently, different internal interaction effects occur and lead to an observed effective inelastic deformation behaviour of such cables. Cyclic bending experiments, Fig. 1 *centre*, show open hysteresis loops with noticeable difference between the first load cycle and the following ones [1, 2], as shown in Fig. 1 *left*. In the framework

D. Manfredo (✉) · V. Dörlich · J. Linn
Fraunhofer ITWM, Kaiserslautern, Germany
e-mail: davide.manfredo@itwm.fraunhofer.de; vanessa.doerlich@itwm.fraunhofer.de; joachim.linn@itwm.fraunhofer.de

M. Arnold
Institute of Mathematics, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
e-mail: martin.arnold@mathematik.uni-halle.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_35

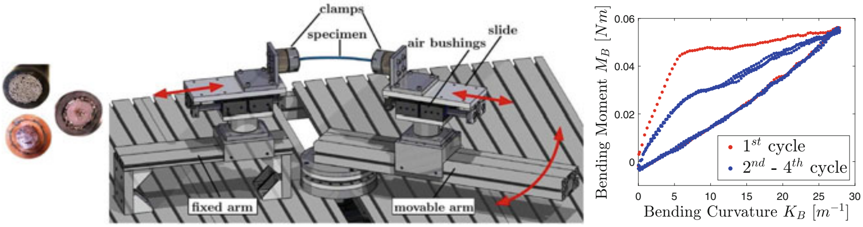


Fig. 1 *Left*: cross sections of different electric cables. *Centre*: pure bending test rig. *Right*: bending moment vs. bending curvature diagram measured in a pure bending experiment

of continuum mechanics, such deformation effects are modelled using suitable constitutive equations for specific material behaviour. In the presented work, we aim at modelling the observed behaviour on an abstract level using hysteresis operators. The choice of this mathematical framework is motivated by the ability of such operators to describe hysteresis phenomena with enough generality and without the need of a priori assumptions on the material behaviour.

2 Hysteresis Operators

As shown in [3, 4], hysteresis operators are a well-studied topic with a variety of applications, mainly hysteresis effects arising from electric and magnetic phenomena. Such operators are normally used to describe the relation between two scalar time-dependent quantities that cannot be expressed in terms of a single-valued function.

2.1 Relay Operator

Given any couple $(a_1, a_2) \in \mathbb{R}^2$ with $a_1 < a_2$, we introduce the relay operator $\mathcal{R}_{a_1,a_2}$. For any input function $v \in C([0, T])$ and initial value $\xi \in \{\pm 1\}$, the output $w = \mathcal{R}_{a_1,a_2}[v] \colon [0, T] \to \{\pm 1\}$ is equal to -1 if the input function value $v(t)$ crosses the threshold a_1 from above, and is equal to $+1$ if $v(t)$ crosses the threshold a_2 from below.

The relay operator can be interpreted as a switch operator between the values -1 and $+1$, with switching interval of width $a_2 - a_1$ and centered in $(a_2 - a_1)/2$. A graphical representation of the relay operator is given in Fig. 2. A formal definition of the relay operator can be found in [4].

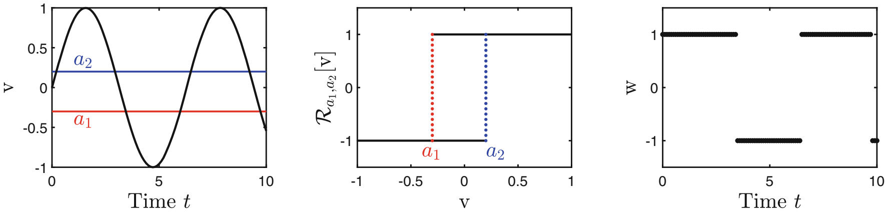


Fig. 2 *Left*: input function $v(t) = \sin(t)$, with $t \in [0, 10]$. *Centre* diagram of the relay operator with $a_1 = -0.3$ and $a_2 = 0.2$. *Right*: output function $w(t) = \mathcal{R}_{a_1,a_2}[v](t)$, with initial value $\xi = +1$

2.2 Preisach Operator

The previously described relay operator is the "building block" of the Preisach operator. To be more precise, a superposition of relay operators multiplied by a suitable kernel function $\omega(r, s)$, assumed to vanish for large values of $|s|$ and r, defines the Preisach operator.

$$w(t) = \mathcal{P}[v](t) = \int_0^{+\infty} \int_{-\infty}^{+\infty} \omega(r, s)\mathcal{R}_{s-r,s+r}[v](t)\, dsdr. \qquad (1)$$

Here, v and w are respectively the input (Fig. 3 *top left*) and the output function, s and r are the coordinates of the Preisach plane, and $\mathcal{R}_{s-r,s+r}$ is the relay operator.

If we consider an input function $v(t)$, for every time t we determine the set

$$A_{\pm}(t) = \{(r, s) \in \mathbb{R}_+ \times \mathbb{R} \colon \mathcal{R}_{s-r,s+r}[v](t) = \pm 1\}.$$

The union of such sets corresponds to the so-called Preisach plane, as will be explained in Sect. 3. One can verify that the dividing line $B(t) = \partial A_+(t) \cap \partial A_-(t)$, also called memory curve, at each time t is the graph of a function which can be defined recursively and carries the total memory information present in the system at time t [3]. In Fig. 3 *top right*, two examples of memory curves are shown. Using $\mathcal{R}_{s-r,s+r}[v](t) \in \{\pm 1\}$ and the definition of $A_{\pm}(t)$, (1) can be rewritten as

$$w(t) = \int_{A_+(t)} \omega(r, s)\, dsdr - \int_{A_-(t)} \omega(r, s)\, dsdr.$$

It should be noted that Preisach hysteresis operators provide a model for causal response [4], such that the output value $w(t)$ at time t depends only on inputs $v(\bar{t})$ at past times $\bar{t} \leq t$. Thus, hysteresis loops can be computed by integrating a suitable kernel function $\omega(r, s)$ over a domain included in the Preisach plane.

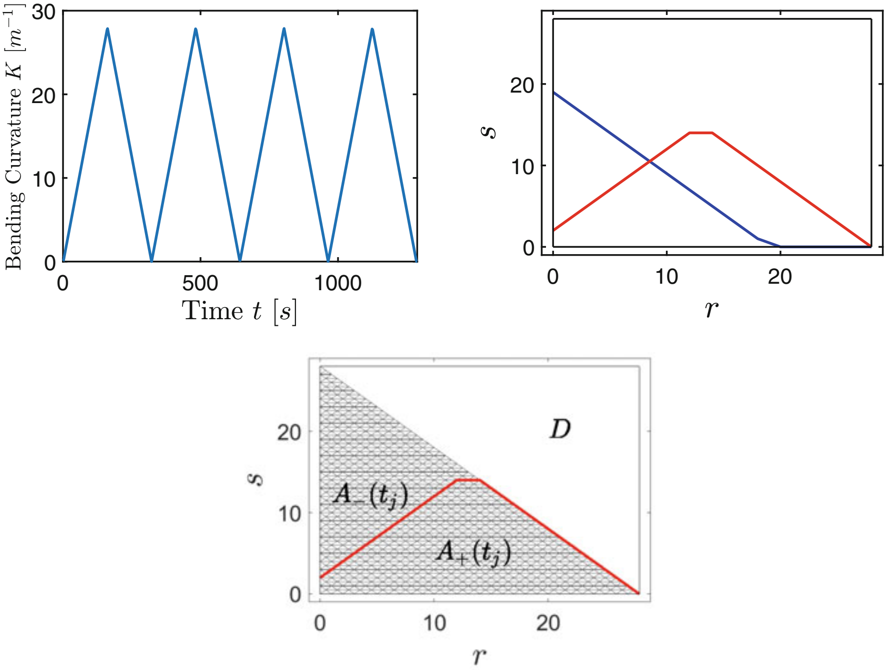


Fig. 3 *Top left*: input given as curvature vs. time. *Top right* domain (black rectangle) included in the Preisach plane with two examples of memory curve. *Bottom*: domain included in the Preisach plane with the triangulation and a memory curve for a given time t_j

3 Problem Formulation

As previously said, we aim at describing the input–output relation of bending curvature vs. bending moment by means of the Preisach operator, utilising data coming from a pure bending cyclic experiment. The available data are time $\{t_i\}_{1\leq i\leq T}$, bending curvature $\{K_i\}_{1\leq i\leq T}$ and bending moment $\{M_i\}_{1\leq i\leq T}$. Note that the values of time and bending curvature are prescribed by the experimental procedure, while the values of bending moment are measured.

Starting from the input function, for each time step t_i, we recursively define the Preisach plane, i.e. the sets $A_\pm(t_i)$ and the memory curve $B(t_i)$. Thus, our goal is to find $\omega(r, s)$ such that the following expression is minimised

$$\frac{1}{T}\sum_{i=1}^{T}\frac{1}{2}\left(M_i - \int\int_{A_+(t)} \omega(r,s)\,dsdr + \int\int_{A_-(t)} \omega(r,s)\,dsdr\right)^2 . \tag{2}$$

To this end, we will take into account only a subset of the Preisach plane, namely the rectangle $[0, \max_{0\leq i\leq T}\{K_i\}] \times [0, \max_{0\leq i\leq T}\{K_i\}]$, since we assume $\omega(r, s)$ to vanish outside such domain. Moreover, as shown in [5], we choose a tolerance d to round the input values. Hence, we divide the part of the Preisach plane crossed by the memory curve $B(t)$ in $n-1$ triangles of equal area, such that at each time step, $B(t_i)$ lies on the edges of the triangles, see Fig. 3 *bottom*. Now, we denote by $X \subset \mathbb{N}_+$ the set of indices given to the elements of the triangulation, by e^m, with $m \in X$, the triangles of the grid, and we define the sets

$$\begin{aligned} X_i &= \{m \in X | e^m \text{ below the memory curve at time } t_i\}, \\ X \backslash X_i &= \{m \in X | e^m \text{ above the memory curve at time } t_i\}. \end{aligned}$$

As shown in Fig. 3 *top left*, we call D the part of the Preisach plane that is never crossed by the memory curve. We assume that the kernel function $\omega(r, s)$ is piecewise constant over each triangle of the mesh and over D, and we want to approximate the output as

$$M_i \approx \sum_{m \in X_i} \int\int_{e^m} \omega(r, s)\, ds dr - \sum_{m \in X \backslash X_i} \int\int_{e^m} \omega(r, s)\, ds dr - c, \quad i = 1, \ldots, T$$

c being the constant value of the kernel function over D. Now, we define the row vector $\mathbf{\Delta}_i = [\delta_i^1, \ldots, \delta_i^{n-1}, -1]$ for each time step t_i, where $\delta_i^m = 1$ if $m \in X_i$ and $\delta_i^m = -1$ if $m \in X \backslash X_i$. Calling $x^m = \int\int_{e^m} \omega(r, s)\, ds dr$, we have

$$\Delta = \begin{bmatrix} \mathbf{\Delta}_1 \\ \vdots \\ \mathbf{\Delta}_T \end{bmatrix} \in \mathbb{R}^{T \times n}, \quad \boldsymbol{X} = \begin{bmatrix} x^1 \\ \vdots \\ x^{n-1} \\ c \end{bmatrix} \in \mathbb{R}^n, \quad \boldsymbol{Y} \in \begin{bmatrix} M_1 \\ \vdots \\ M_T \end{bmatrix} \in \mathbb{R}^T. \tag{3}$$

Hence, using (2) and (3), the function to minimise is $f(\boldsymbol{X}) = \frac{1}{2}\|\Delta \cdot \boldsymbol{X} - \boldsymbol{Y}\|^2$. In practice, one often deals with insufficient experimental data, yielding $\text{rank}(\Delta) = q < \min\{T, n\}$ for the matrix Δ. In order to compensate for the lack of data, we perform a singular value decomposition of the matrix $\Delta^T \Delta = USV^T$, where S is a diagonal matrix, with $\text{rank}(S) = q$.

We extract $\hat{S}$, $\hat{U}$, $\hat{V}$ from S, U, V, respectively, by eliminating the rows and the columns of S that are zero, and the corresponding columns of U and V. Setting $\boldsymbol{X} = \hat{V}\boldsymbol{Z}$, the expression to minimise becomes $g(\boldsymbol{Z}) = \boldsymbol{Z}^T \hat{S} \boldsymbol{Z} - \boldsymbol{Y}^T \Delta \cdot \hat{V} \boldsymbol{Y}$. It is easily verified, that once a minimiser $\boldsymbol{Z}^*$ of g is found, then $\boldsymbol{X}^* = \hat{V}\boldsymbol{Z}^*$ minimises f.

4 First Results and Conclusion

A minimiser $\boldsymbol{Z}^*$ of g can be found using a Matlab routine such as “quadprog”. In Fig. 4 *left*, an approximation of the kernel function $\omega(r, s)$ is shown, and the integral

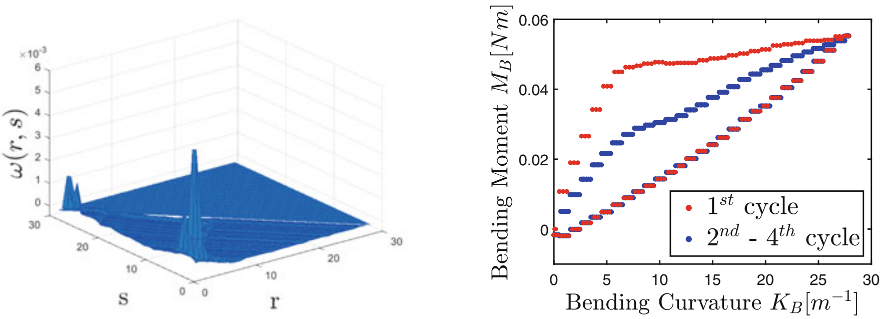


Fig. 4 *Left*: kernel function obtained by the minimisation of g. *Right*: estimated plot of bending moment vs. curvature obtained by means of the hysteresis operator

of such kernel function over the domain included in the Preisach plane results in the diagram shown in Fig. 4 *right*. Comparing the experimental data in Fig. 1 *right* with the diagram in Fig. 4 *right*, one can see that this approach describes the input–output relation as bending curvature vs. bending moment observed during the experiments quite well. One should note that the step-like behaviour of the diagram in Fig. 4 *right* is due to the tolerance value d. However, the kernel function shows a highly nonlinear behaviour, and further work is necessary to investigate if its shape and properties are related to the physics of the studied phenomenon.

The Preisach operator is a very powerful and versatile tool to describe inelastic deformation behaviours of electric cables and the consequent open hysteresis loops arising from bending experiments. Moreover, such a mathematical tool captures the difference between load cycles very well and is relatively easy to implement. A more detailed study of the properties of the kernel function is necessary, with particular focus on its relation with the experimental data and the physics of the phenomenon.

Acknowledgments This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860124.

References

1. V. Dörlich, J. Linn, S. Diebels, Flexible Beam-Like Structures - Experimental Investigation and Modeling of Cables, in H. Altenbach et al., Advances in Mechanics of Materials and Structural Analysis. Advanced Structured Materials, vol 80, Springer International Publishing, 2018, pp. 27–46.
2. V. Dörlich, J. Linn, S. Diebels, Bending of Viscoplastic Cables, Proc. Appl. Math. Mech. 17 (2017), 293–294.
3. M. Brokate, J. Sprekels, Hysteresis and Phase Transitions, Springer-Verlag, New York, 1996.
4. A. Visintin, Differential Models of Hysteresis, Springer-Verlag, Berlin, Heidelberg, 1994, vii+407 pp.
5. M. E. Shirley, R. Venkataraman, On the Identification of Preisach Measures. Proceedings Volume 5049, Smart Structures and Materials 2003: Modeling, Signal Processing, and Control.

High-Throughput Analysis of Potato Vitality

Elisa Atza and Neil Budko

Abstract Vitality is a fundamental trait for the development of a plant. It is known to depend on various factors, such as climate, soil, and the plant's genetics, but the progressive depletion of soil nutrients make it a priority for the industry to pinpoint which of the controllable qualities of a seed have the biggest impact on vitality. This work describes techniques applied in a high-throughput phenotyping project, the first of this magnitude for a complex plant, the potato (*solanum tuberosum*). We also present the results of an analysis of associations between the chemical composition of the seed potatoes and field performance, solving the arising underdetermined linear systems by means of PLS regression. We show that some but not all of the chemical data is strongly associated to vitality.

1 Introduction

A potato plant is vital if it manifests in a large canopy and exhibits homogeneous growth in the early stages of its development. Potato seed producers as well as farmers have noticed that potato seeds of the same cultivar perform differently in the same conditions depending on the field in which the seed tubers have been produced.

A cultivar, or variety, is described as a set of plants for which specific characteristics are reliably passed on to the offspring. Uniform growth facilitates farming thus high variability in the growth of a variety is undesirable.

In collaboration with potato seed producers HZPC and Averis seeds, we aim to quantify the contribution of non genetic factors to the plant development. Specifically, we investigate the link between the chemical and biological properties of a tuber, and the vitality of the sprouting plant. Identifying relevant markers would allow for a screening of tubers prior to planting; allowing for higher yields and customised offers to the clients. There is no standard way to relate such a broad

E. Atza (✉) · N. Budko
Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
e-mail: e.atza@tudelft.nl; n.v.budko@tudelft.nl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_36

variety of interdependent data regarding a tuber to the measured development of the plant.

In order to model this problem we study six different cultivars, and for each cultivar we measure 30 different tubers, which are genetically identical, but have either been produced in different locations or have received a specific treatment. These 30 different tubers we call batches, so that in total we study 180 different batches belonging to 6 varieties.

For the experiment, data was collected from the studied tubers before and after planting over several consecutive years. Our industrial partner, HZPC, is responsible for the collection of most of the tuber data, as well as for the planting process. Aerial pictures are collected for the field experiment in different European locations by a commercial drone operator, which provides us with orthophotos of the fields according to industrial standard. We then process these aerial pictures ourselves in order to quantify vitality from expressed traits of the plant, a process referred to as phenotyping.

We will shortly present the procedure used to extract canopy coverage in the field from drone images, and then discuss the first associations resulting from linear regression performed considering the different data sets as independent variables.

2 Linear Regression with PLS

We predict vitality parameters $Y \in \mathbb{R}^{180}$, from different tuber data $X \in \mathbb{R}^{180\times p}$, by investigating the presence of a linear dependence:

$$Y = X\beta + \epsilon. \tag{1}$$

2.1 Response

The experiment fields are planted according to a randomized block design, so that four replicates of the 180 batches are distributed on separate non-adjacent parts of the field. The first step in the processing of aerial pictures is to delimit the regions, called plots, where each batch is planted.

For each field we choose one image dated around 35–40 days after planting to find these boundaries. At this point in their development, plants within one plot form a continuous canopy and simultaneously have not grown enough to bridge the gap to the next plot. Thus, looking for gaps in the vegetation at this stage almost coincides with looking for plot boundaries.

We use both physical markers on the field and manual input to determine the region of interest in the drone image, then algorithmically look for gaps inside this region. Knowing the number of plots and the number of columns (*ridges*) in each portion of the field, our algorithm determines the most likely plot boundaries.

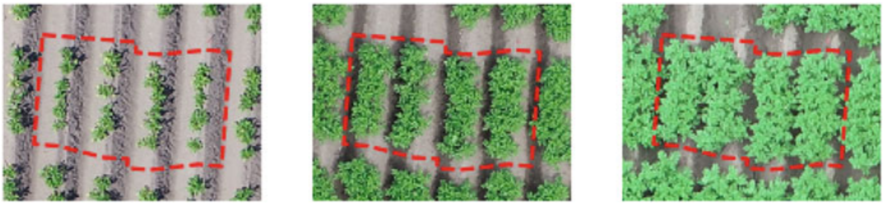


Fig. 1 Plot boundaries are detected in the middle image, these boundaries are then used on other photos after time alignment. Each plot is subdivided in four columns, called ridges

After having determined and visually inspected the plot boundaries found on this date, we use physical marks present on the field to align all photos of the same field, such that the boundaries found can be used both before and after the reference date, when canopies' growth makes it harder to distinguish plots, or when dealing with delayed sprouting and small canopies. An example is given in Fig. 1.

Given the resolution of the orthophotos, we look at four vitality measurements per plot, namely we quantify the mean canopy coverage in each ridge. In this way we obtain 16 canopy measurements (four ridges times four plots) for each of the 180 batches on any of the r measurement dates, i.e. our response $\mathbf{Y} \in \mathbb{R}^{2880 \times r}$.

The mean ridge canopy data must be corrected for possible smooth spatial variations across the field due to large-scale inhomogeneities in soil properties and other factors influencing the growth of plants.

For each measurement date $j = 1, \ldots, r$ we model the spatial variations in each column $\mathbf{Y}(j)$ as

$$\mathbf{Y}(j) = X_1 \beta_1 + \boldsymbol{\epsilon}_1, \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

where $\beta_1 = [c_1, c_2, c_3, c_4]^T \in \mathbb{R}^{p_1}$, $p_1 = 4$. The structure of the design matrix $X_1 \in \mathbb{R}^{n \times p_1}$, $n = 2880$, can be inferred from (3), which is the row-wise expression of (2), and σ^2 is the field specific variance, which we estimate from the data.

For a single ridge i, $i = 1, \ldots, 2880$, located at pixel coordinates $\langle x_i, y_i \rangle$ the model in (2) translates to:

$$Y_i = c_1 + c_2 x_i + c_3 y_i + c_4 x_i y_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

For spatial correction we retain the field mean, but the global linear and bi-linear spatial variations are removed:

$$\mathbf{Y}_{\text{corr}}(j) = \mathbf{Y}(j) - X_1 \hat{\beta}_1 + \hat{c}_1 \mathbf{1} = \hat{\boldsymbol{\epsilon}}_1 + \hat{c}_1 \mathbf{1}. \tag{4}$$

where $\hat{\beta}_1$ is the restricted maximum likelihood (REML) estimate of β_1 and $\hat{\boldsymbol{\epsilon}}_1 \sim \mathcal{N}(0, \hat{\sigma}^2 I)$, where $\hat{\sigma}^2$ is the REML estimate of σ^2.

After correction we consider the average growth performance of a batch over multiple days and multiple repetitions reducing the size of our response to $Y \in$

$\mathbb{R}^{180\times 1}$. This is then normalized to have zero mean and unit standard deviation and is our response for the model in (1).

2.2 Predictors

Several aspects of the tubers are analyzed in the scope of the project with the goal of obtaining an exhaustive description of the chemical and biological profile of different batches of the same variety.

In this work we look at three tuber related datasets and we will compare their performance as predictors of vitality:

- Fourier transform infrared (**FTIR**) spectroscopy: for each sample we obtain a spectrum, i.e. a discretized curve, whose values are the absorbances of the sample for given wavenumbers, in this case the matrix X is of size 180×2388.
- Hyperspectral imaging (**HSI**): each sample is photographed at l different wavelenghts resulting in l images of size $w \times h$. This results in an array of size $w \times h \times l$. The values of a pixel at different wavelengths form an array of length l. Averaging these arrays over particular regions of the tuber we obtain spectra for known tuber compartments, such as pith and cortex. In this case $l = 288$, thus we obtain for each compartment a matrix of predictors X of size 180×288.
- X-Ray fluorescence (**XRF**): this technique gives us concentrations of 10 chemical elements in the samples, in this case the predictor matrix X has size 180×10.

Also for our predictors we apply a zero mean and unit standard deviation normalization. Additionally, for the spectral data (FTIR, HSI) we explore normalization by applying the Savitzky-Golay first polynomial derivative (SG1).

For two data sets (FTIR, and HSI) the linear model in (1) is highly underdetermined. We use partial least squares (PLS) regression to solve the resulting system of equations.

2.3 Method

PLS, also called projection to latent structures, is a dimensionality reduction technique for which the explanatory and the dependent variables are both projected on new *components* constructed to maximize the covariance between X and Y, see [1] and [2]. The decomposition of both matrices X and Y is given by the following:

$$X = TP^T + E, \quad T = (\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k), \tag{5}$$

$$Y = UQ^T + F, \quad U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k). \tag{6}$$

Here, T, U contain the k latent vectors as columns. Matrices P, Q, are the matrices of loadings, E and F are the residuals.

The columns of the matrices T and U, the latent vectors, are constructed iteratively by finding weights w_i and c_i for which $t_i = Xw_i$ and $u_i = Yc_i$ with the constraints that $w_i^T w_i = 1$, $t_i^T t_i = 1$ and such that $t_i^T u_i$, which is proportional to the covariance of t_i and u_i, is maximal. Each subsequent column is constructed to be orthogonal to the previous ones, and lastly the matrix T of latent vectors for X is used to predict Y with ordinary least squares (OLS). The maximum number of components of the matrix T is equal to the rank of X, at which point the PLS estimator for the coefficients β will be equal to the minimum length least square estimator, [3], which will have large variance for highly collinear spectroscopic data, [4], thus choosing the number of components to be used is a critical point in the application of this method.

As is usual we split our data in train and test set, in order to find the appropriate number of PLS components, we train models with an increasing number of components up to a preset maximum and at each iteration we use k-fold stratified cross validation, $k = 10$, to evaluate the mean squared error, MSE. We choose then the number of components for which the mean of the k MSEs was minimal. Then we train a model with the optimal number of components, which we evaluate on the test set using the coefficient of determination, R^2, and MSE.

This splitting and training is repeated multiple times, the scores R^2 and MSE are stored in the vectors $\mathbf{R^2}$, and $\mathbf{MSE}$ respectively, so that we can test the robustness of our model by making sure that the empirical standard deviations of the vectors $\sigma(\mathbf{R^2})$ and $\sigma(\mathbf{MSE})$ have a sufficiently small value.

For the XRF dataset we estimate the regression coefficients with OLS.

3 Results

All data in Tables 1, 2, 3, and 4 is displayed in ascending order with respect to the mean R^2. The regression scores are presented for each field and for each year separately. In the case of spectroscopic data we present the results obtained for different normalizations of the tuber data on separate lines.

From our analysis we notice a strong association of the FTIR data set to vitality, regardless of the applied normalization, our evaluation parameters stay consistent for each field.

The regression on FTIR and HSI spectra shows that the prediction performance is influenced by the field in which the vitality has been measured. Furthermore we see that XRF as a stand-alone dataset is not a sufficiently good predictor of vitality, and that the subdivision of hyperspectral data in separate tuber compartments does not offer a substantial difference in performance.

Table 1 Regression on HSP data for year 2

Field	Part	Normal.	# comp	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	pith	STD	38	0.27	0.05	0.27	0.02
C	pith	SG1	39	0.27	0.05	0.27	0.02
C	cortex	SG1	39	0.30	0.04	0.26	0.01
C	cortex	STD	39	0.30	0.04	0.26	0.01
B	pith	STD	33	0.38	0.06	0.39	0.04
B	pith	SG1	37	0.38	0.06	0.39	0.04
B	cortex	SG1	39	0.38	0.06	0.38	0.04
B	cortex	STD	39	0.38	0.06	0.38	0.04
A	pith	SG1	37	0.44	0.08	0.40	0.06
A	pith	STD	38	0.44	0.08	0.40	0.05
A	cortex	STD	39	0.48	0.06	0.38	0.05
A	cortex	SG1	39	0.48	0.06	0.38	0.05

Table 2 Regression on XRF data for both years

Field	Year	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	2	0.21	0.03	0.32	0.04
B	2	0.24	0.05	0.37	0.03
C	1	0.33	0.04	0.36	0.01
A	2	0.33	0.04	0.32	0.02
B	1	0.42	0.02	0.39	0.03
A	1	0.44	0.01	0.31	0.02

Table 3 Regression on FTIR data for year 1

Field	Normal.	# comp.	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	STD	28	0.67	0.02	0.17	0.01
C	SG1	30	0.67	0.02	0.17	0.01
B	STD	19	0.81	0.01	0.12	0.01
B	SG1	21	0.81	0.01	0.12	0.01
A	STD	30	0.84	0.02	0.09	0.01
A	SG1	29	0.84	0.02	0.09	0.01

Table 4 Regression on FTIR data for year 2

Field	Normal.	# comp.	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	SG1	14	0.62	0.02	0.14	0.01
C	STD	14	0.62	0.03	0.14	0.01
B	SG1	22	0.80	0.01	0.13	0.01
B	STD	22	0.80	0.01	0.13	0.01
A	STD	23	0.84	0.02	0.11	0.01
A	SG1	24	0.84	0.02	0.11	0.01

4 Conclusions and Further Research

FTIR data is the best performing, and most consistent in predictive power over different years. Ongoing research suggests that a more tailored analysis of HSI data could improve its predictive performance. Furthermore, the strong link between prediction performance and field of measurement, as well as the fitness of non-linear models for regression on chemical datasets should be investigated.

References

1. P.H. Garthwaite, An Interpretation of Partial Least Squares, Journal of the American Statistical Association 89, No. 425 (1994), 122–27,
2. M. Haenlein and A.M. Kaplan, A Beginner's Guide to Partial Least Squares Analysis, Understanding Statistics, 3:4 (2004), 283–297.
3. A. Phatak and F. de Hoog, Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS, J. Chemometrics, 16 (2002), 361–367.
4. S. Wold, M. Sjöström, and L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, Volume 58, Issue 2, 2001, Pages 109–130,

Optimized Hydrodynamical Model for Charge Transport in Graphene

Vito Dario Camiola, Giovanni Nastasi, Vittorio Romano, and Giorgia Vitanza

Abstract Starting from the Boltzmann equations and employing the moment method, hydrodynamical models for charge transport in suspended monolayer graphene have been devised. In particular in Camiola and Romano (J Stat Phys 157:1114–1137, 2014), Luca and Romano (Ann Phys 406:30–53, 2019), Luca and Romano (Int J Non-Linear Mech 104:39–58, 2018), Luca and Romano (Ann Phys 406:30–53, 2019), Luca et al. (J Comput Theoret Trans 49(7), 2020), and Camiola et al. (Charge transport in low dimensional semiconductor structures, the maximum entropy approach. Springer, 2020) closure relations have been obtained by adopting the Maximum Entropy Principle (MEP). Stemming from the kinetic equations, some physical parameters appear in the production terms such as the acoustic phonon, the optical phonon and the K-phonon coupling constants. Their values have been estimated by experimental data and fundamental approach, e.g. the density functional theory. However, they depend on the modelling of the energy band and scattering terms. Here, we try to improve the hydrodynamical model proposed in Camiola and Romano (J Stat Phys 157:1114–1137, 2014) by an optimisation of the parameters above through a minimisation of the difference between velocity and energy, found with the considered hydrodynamical models and the direct solution of the Boltzmann equation obtained with a Discontinuous Galerkin (DG) method (Coco et al., Ricerche mat 66:201–220, 2017; Majorana et al., Commun Comput Phys 26, 114–134, 2019).

V. D. Camiola · G. Nastasi (✉) · V. Romano · G. Vitanza
University of Catania, Catania, Italy
e-mail: dario.camiola@unict.it; g.nastasi@unict.it; romano@dmi.unict.it; giorgia.vitanza@phd.unict.it

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_37

1 Boltzmann Equation

In a semiclassical kinetic setting, the charge transport in graphene is described, in general, by four Boltzmann equations, one for electrons in the valence (π) band and one for electrons in the conduction (π^*) band, that in turn can belong to the K or K' valley. Here, we assume that the K and K' valleys are equivalent. Moreover, by applying a gate voltage transversal with respect to the graphene sheet, it is possible to modify the Fermi energy ε_F and therefore the charge density. As shown in [10], if the Fermi energy is high enough (more than about 0.2 eV), the contribution to the current due to holes in the valence band is negligible with respect to that of electrons in the conduction band. Therefore, only the transport equation for electrons in the conduction band is considered and interband transitions are neglected. It can be written as

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f - \frac{e}{\hbar} \mathbf{E} \cdot \nabla_{\mathbf{k}} f = C(\mathbf{k}), \tag{1}$$

where $f = f(t, \mathbf{x}, \mathbf{k})$ represents the distribution function of electrons in the conduction band at position $\mathbf{x}$, time t and wave-vector $\mathbf{k}$. We denote by $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{k}}$ the gradients with respect to the position and the wave vector, respectively. The group velocity $\mathbf{v}$ is related to the energy band ϵ by $\mathbf{v} = \frac{1}{\hbar} \nabla_{\mathbf{k}} \epsilon$. With a very good approximation [3], a linear dispersion relation holds for the energy bands around the Dirac points; so that, choosing the origin of the reference frame in the $\mathbf{k}$-space coinciding with a Dirac point, we have $\epsilon = \hbar \mathbf{v}_F |\mathbf{k}|$, where $\mathbf{v}_F$ is the (constant) Fermi velocity and $\hbar$ the Planck constant divided by 2π.

The Brillouin zone is extended to $\mathbb{R}^2$. The elementary (positive) charge is denoted by e. Here the electric field $\mathbf{E}$ is assumed as external, and therefore we do not include the Poisson equation. The right-hand side of Eq. (1) is the collision term which takes into account scatterings between electrons and phonons. In suspended monolayer graphene three kinds of phonons have to be considered: acoustic, optical and K phonons. We assume that phonons are in thermal equilibrium.

The collision term can be written as

$$C(\mathbf{k}) = \int_{\mathbb{R}^2} S(\mathbf{k}', \mathbf{k}) f(t, \mathbf{x}, \mathbf{k}')(1 - f(t, \mathbf{x}, \mathbf{k}))\, d\mathbf{k}' \\ - \int_{\mathbb{R}^2} S(\mathbf{k}, \mathbf{k}') f(t, \mathbf{x}, \mathbf{k})(1 - f(t, \mathbf{x}, \mathbf{k}'))\, d\mathbf{k}' \tag{2}$$

where the total transition rate is given by the sum of the contributions of the above mentioned types of scatterings

$$S(\mathbf{k}', \mathbf{k}) = \sum_{\nu} |G^{(\nu)}(\mathbf{k}', \mathbf{k})|^2 \Big[\left(n_{\mathbf{q}}^{(\nu)} + 1 \right) \delta \left(\epsilon(\mathbf{k}) - \epsilon(\mathbf{k}') + \hbar\omega_{\mathbf{q}}^{(\nu)} \right) \\ + n_{\mathbf{q}}^{(\nu)} \delta \left(\epsilon(\mathbf{k}) - \epsilon(\mathbf{k}') - \hbar\omega_{\mathbf{q}}^{(\nu)} \right) \Big]$$

The index ν labels the ν-th phonon mode, $|G^{(\nu)}(\mathbf{k}', \mathbf{k})|$ is the matrix element, which describes the scattering mechanism, due to phonons of type ν. The symbol δ denotes the Dirac distribution, $\omega_{\mathbf{q}}^{(\nu)}$ is the ν-th phonon frequency, $n_{\mathbf{q}}^{(\nu)}$ is the Bose-Einstein distribution for the phonons of type ν

$$n_{\mathbf{q}}^{(\nu)} = \frac{1}{e^{\hbar\omega_{\mathbf{q}}^{(\nu)}/k_B T} - 1},$$

k_B is the Boltzmann constant and T is the graphene lattice temperature which, in this article, will be assumed constant.

For acoustic phonons, one usually considers the elastic approximation [5]

$$2n_{\mathbf{q}}^{(ac)} |G^{(ac)}(\mathbf{k}', \mathbf{k})|^2 = \frac{1}{(2\pi)^2} \frac{\pi D_{ac}^2 k_B T}{2\hbar \sigma_m v_p^2} (1 + \cos\vartheta_{\mathbf{k},\mathbf{k}'}), \tag{3}$$

where D_{ac} is the acoustic phonon coupling constant, v_p is the sound speed in graphene, σ_m the graphene areal density, and $\vartheta_{\mathbf{k},\mathbf{k}'}$ is the convex angle between $\mathbf{k}$ and $\mathbf{k}'$.

There are three relevant optical phonon scatterings: the longitudinal optical (LO), the transversal optical (TO) and the K-phonons. The matrix elements are [12]

$$|G^{(LO)}(\mathbf{k}', \mathbf{k})|^2 + |G^{(TO)}(\mathbf{k}', \mathbf{k})|^2 = \frac{2}{(2\pi)^2} \frac{\pi D_O^2}{\sigma_m \omega_O}, \tag{4}$$

$$|G^{(K)}(\mathbf{k}', \mathbf{k})|^2 = \frac{2}{(2\pi)^2} \frac{\pi D_K^2}{\sigma_m \omega_K} (1 - \cos\theta_{\mathbf{k},\mathbf{k}'}), \tag{5}$$

where D_O is the optical phonon coupling constant, ω_O is the optical phonon frequency, D_K is the K-phonon coupling constant, and ω_K is the K-phonon frequency.

2 Hydrodynamical Model: L6MM

We will investigate the model proposed in [1, 2, 6–8] (for quantum corrections see also [9]) which is based on the following moments

$$\rho = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(t, \mathbf{x}, \mathbf{k}) d^2\mathbf{k} \quad \text{density},$$

$$\rho W = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(t, \mathbf{x}, \mathbf{k}) \epsilon(\mathbf{k}) d^2\mathbf{k} \quad \text{energy density,}$$
$$\rho \mathbf{V} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(t, \mathbf{x}, \mathbf{k}) \mathbf{v}(\mathbf{k}) d^2\mathbf{k} \quad \text{linear momentum density,}$$
$$\rho \mathbf{S} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} f(t, \mathbf{x}, \mathbf{k}) \epsilon(\mathbf{k}) \mathbf{v}(\mathbf{k}) d^2\mathbf{k} \quad \text{energy-flux density.}$$

The corresponding evolution equations are given by

$$\frac{\partial}{\partial t}\rho + \nabla_{\mathbf{x}}(\rho \mathbf{V}) = 0,$$
$$\frac{\partial}{\partial t}(\rho W) + \nabla_{\mathbf{x}}(\rho \mathbf{S}) + e\rho \mathbf{E} \cdot \mathbf{V} = \rho C_W,$$
$$\frac{\partial}{\partial t}(\rho \mathbf{V}) + \nabla_{\mathbf{x}}(\rho \mathbf{F}^{(\mathbf{0})}) + e\rho \mathbf{G}^{(\mathbf{0})} : \mathbf{E} = \rho C_{\mathbf{V}},$$
$$\frac{\partial}{\partial t}(\rho \mathbf{S}) + \nabla_{\mathbf{x}}(\rho \mathbf{F}^{(\mathbf{1})}) + e\rho \mathbf{G}^{(\mathbf{1})} : \mathbf{E} = \rho C_{\mathbf{S}}.$$

Besides the average densities, velocities, energies and energy fluxes, additional quantities appear

$$\rho C_{\mathbf{V}} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \mathbf{v}(\mathbf{k}) C(\mathbf{k}) d^2\mathbf{k}, \tag{6}$$
$$\rho C_W = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \epsilon(\mathbf{k}) C(\mathbf{k}) d^2\mathbf{k}, \tag{7}$$
$$\rho C_{\mathbf{S}} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \epsilon(\mathbf{k}) \mathbf{v}(\mathbf{k}) C(\mathbf{k}) d^2\mathbf{k}, \tag{8}$$
$$\rho \begin{pmatrix} \mathbf{F}^{(\mathbf{0})} \\ \mathbf{F}^{(\mathbf{1})} \end{pmatrix} = \frac{2}{(2\pi)^2} \int_{\mathbb{R}^2} \begin{pmatrix} 1 \\ \epsilon(\mathbf{k}) \end{pmatrix} \mathbf{v}(\mathbf{k}) \otimes \mathbf{v}(\mathbf{k}) f(t, \mathbf{x}, \mathbf{k})\, d^2\mathbf{k}, \tag{9}$$
$$\rho \begin{pmatrix} \mathbf{G}^{(\mathbf{0})} \\ \mathbf{G}^{(\mathbf{1})} \end{pmatrix} = \frac{2}{\hbar(2\pi)^2} \int_{\mathbb{R}^2} f(t, \mathbf{x}, \mathbf{k}) \nabla_{\mathbf{k}} \begin{pmatrix} \mathbf{v}(\mathbf{k}) \\ \epsilon(\mathbf{k}) \mathbf{v}(\mathbf{k}) \end{pmatrix} d^2\mathbf{k}, \tag{10}$$

that must be expressed as functions of the basic variables ρ, W, $\mathbf{V}$, $\mathbf{S}$. Regarding the production terms, they are given by the sum of contributions arising from the different types of phonon scattering

$$C_M = C_M^{(ac)} + \sum_{\nu = LO, TO, K} C_M^{(\nu)}$$

with $M = \rho, W, \mathbf{V}, \mathbf{S}$. We recall that the generic term due to a single scattering from a state $\mathbf{k}$ to a state $\mathbf{k}'$ is given by (2). Explicit closure relations have been obtained in

[2] by adopting MEP and by linearising the resulting distribution f_{MEP} with respect to the vectorial Lagrange multipliers. We will refer to this model as L6MM.

3 Formulation of the Problem

By inserting f_{MEP} in the definition of the quantities appearing in L6MM one gets a closed system of hyperbolic balance equations. In particular, the production terms contain the electron-phonon coupling parameters D_{ac}, D_{Γ}, D_K. We try to improve the accuracy of L6MM with respect to the mean values of velocity and energy obtained by a direct solution of the Boltzmann equation, considering D_{ac}, D_{Γ}, D_K as fitting parameters which are allowed to vary with respect to the values present in the Boltzmann equation.

So, in Eqs. (7)–(9), instead of D_{ac}, D_{Γ}^2, D_K^2, we consider $a_1\, D_{ac}$, $a_2\, D_{\Gamma}^2$, $a_3\, D_K^2$ where the coefficients a_i belong to a suitable admissible set we specify below. In the case $a_1 = a_2 = a_3 = 1$ one has the value used in the Boltzmann equation which will be assumed as an initial guess a in the optimisation procedure.

Several 1D space homogeneous solutions with different values of the only significant component of the electric field E and Fermi energy ε_F have been considered. In this case, the density is constant and depends on ε_F (see [4, 10]). The steady-state solutions are compared.

We take the following objective function

$$f_{\text{obj}}(\mathbf{a}) = \alpha\Big[\sum_{i,j}\big|V_{L6MM}(\mathbf{a}, E_i, \varepsilon_{F_j}, t) - V_{DG}\big|^2\Big]^{1/2} + \beta\Big[\sum_{i,j}\big|W_{L6MM}(\mathbf{a}, E_i, \varepsilon_{F_j}, t) - W_{DG}\big|^2\Big]^{1/2}, \tag{11}$$

where $V_{L6MM}(\mathbf{a}, E_i, \varepsilon_{F_j}, t)$ and $W_{L6MM}(\mathbf{a}, E_i, \varepsilon_{F_j}, t)$ are velocity and energy at time t, electric field E_i and Fermi level ε_{F_j} computed with L6MM, respectively; these functions depend on $\mathbf{a} = [a_1, a_2, a_3]$. V_{DG} and W_{DG} are the reference values calculated with the DG method.

In order to consider the steady state, we have fixed the final time $t = 3$ ps. Moreover, we have set $\alpha = 1$ and $\beta = 0.1$ to give more weight to the velocity with the aim to get an improvement of the current. The parameters a_i are allowed to vary in the range $[0.4 - 2.5]$.

The complete formulation of the problem is as follows:

$$\begin{cases} \min f_{\text{obj}}(\mathbf{a}) \\ \mathbf{a}_0 = [1, 1, 1] \\ 0.4 \le a_k \le 2.5 \qquad k = 1, 2, 3. \end{cases} \tag{12}$$

Ten values of the electric field have been considered, from 0.1 V/μm to 1 V/μm with increments of 0.1 V/μm, and three Fermi levels, 0.4, 0.5 and 0.6 eV.

To solve this constrained optimization problem we have adopted three approaches: the MATLAB optimization function `fmincon` [11], a genetic algorithm, and the simulated annealing method.

4 Numerical Results

In this section, we show the numerical results and highlight the difference between three models: L6MM, DG and the Hydrodynamical model optimized with the new constants $a_1 D_{ac}$, $a_2 D_\Gamma^2$ and $a_3 D_K^2$. The value of the objective function in the initial guess $\mathbf{a}_0$ is $f_{\text{obj}}(\mathbf{a}_0) = 2.819$.

By using the MATLAB function `fmincon`, the optimum is given by the vector $\mathbf{a} = [0.400, 2.115, 0.400]$ with $f_{\text{obj}}(\mathbf{a}) = 2.067$. The genetic algorithm with the number of maximum generation set equal to 100, gives as optimum point $\mathbf{a} = [0.399, 2.088, 0.399]$ with $f_{\text{obj}}(\mathbf{a}) = 2.082$, while the simulated annealing furnishes the optimal solution $\mathbf{a} = [0.401, 2.093, 0.434]$, with $f_{\text{obj}}(\mathbf{a}) = 2.084$. The numerical results in the different cases are very similar, but the genetic algorithm and simulated annealing are more expensive computationally than the function `fmincon`.

In Fig. 1 the transient solutions obtained with the direct solution of the Boltzmann equation by using the DG method, the original L6MM model and the optimized L6MM model are compared. We get a noticeable improvement of the asymptotic value of the velocity, at the expenses of a slight worsening of the asymptotic values of the energy (note that the scales are different between velocity and energy). However, from the point of view of the steady electric current, the overall performance of the improved L6MM is better than the original L6MM.

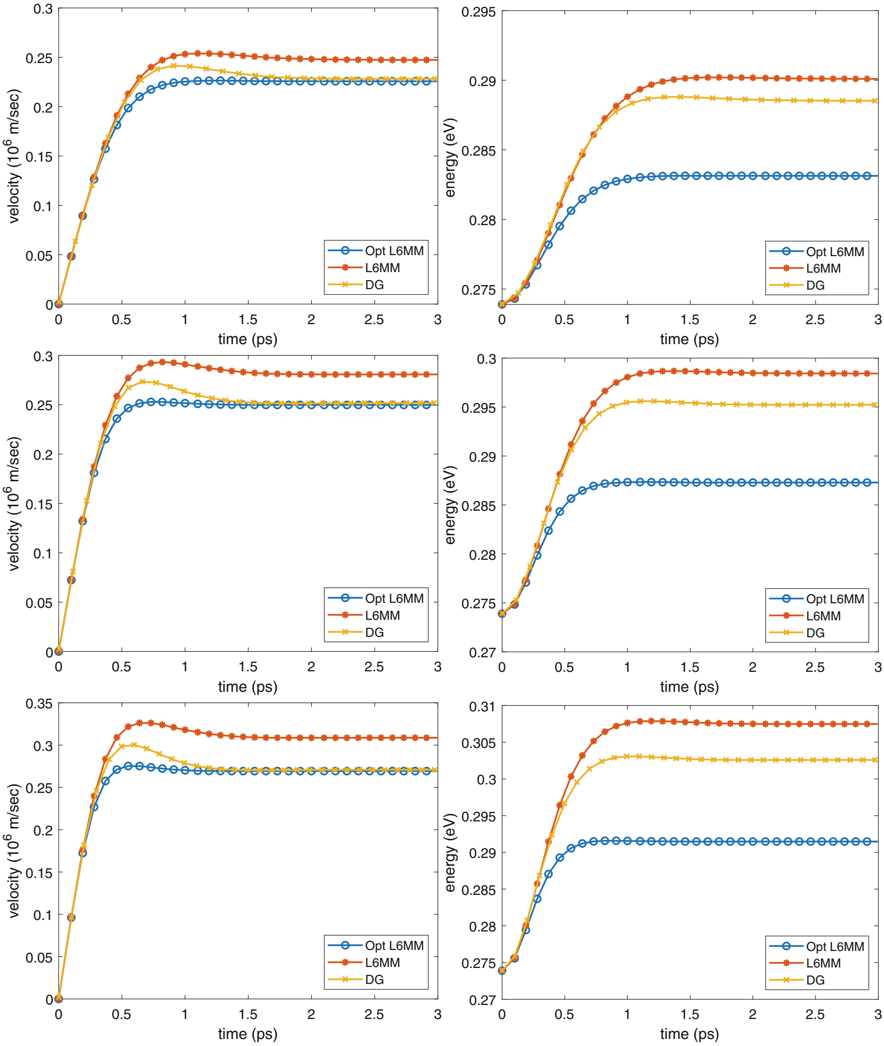


Fig. 1 Velocity (on the left) and Energy (on the right) of charges in graphene, calculated by DG, L6MM and compared with the optimized L6MM (blue line) in the cases $\varepsilon_F = 0.4\,\text{eV}$ and $E = 0.2\,\text{V}/\mu\text{m}$ (top), $E = 0.3\,\text{V}/\mu\text{m}$ (middle), $E = 0.4\,\text{V}/\mu\text{m}$ (bottom)

References

1. Camiola, V.D., Mascali, G., Romano, V.: “Charge Transport in Low Dimensional Semiconductor Structures, The Maximum Entropy Approach”. Springer, (2020).
2. Camiola, V.D., Romano, V.: “ Hydrodynamical Model for Charge Transport in Graphene”. J. Stat. Phys. 157, 1114–1137 (2014).

3. Castro Neto, A. H., Geim, A. K., Guinea, F., Novoselov, K. S., Peres, N. M. R.: “The electronic properties of graphene”, Rev. Mod. Phys., Vol. 81, pp. 109–162 (2009).
4. Coco, M., Majorana, A., Romano, V.: “Cross validation of discontinuous Galerkin method and Monte Carlo simulations of charge transport in graphene on substrate”. Ricerche mat 66, 201–220 (2017)
5. Hwang, E. H., Das Sarma, S.: “Acoustic phonon scattering limited carrier mobility in two-dimensional extrinsic graphene”, Phys. Rev. B, Volume 77, 115449 (2008).
6. Luca, L., Mascali, G., Nastasi, G., Romano, V.: “Comparing Kinetic and MEP Model of Charge Transport in Graphene”, J. Comput. Theoret. Trans., Volume 49, Issue 7 (2020)
7. Luca, L., Romano, V.: “Hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle: the case of moments based on energy powers”. Atti della Accademia Peloritana dei Pericolanti - Classe di Scienze Fisiche, Matematiche e Naturali, [S.l.], p. A5, (2018)
8. Luca, L., Romano, V.: “Comparing linear and nonlinear hydrodynamical models for charge transport in graphene based on the Maximum Entropy Principle”. Int. J. Non-Linear Mech., Volume 104, pp. 39–58 (2018)
9. Luca, L., Romano, V.: “Quantum corrected hydrodynamic models for charge transport in graphene”. Annals of Physics, Volume 406, pp. 30–53 (2019)
10. Majorana, A., Nastasi, G., Romano,V.: “Simulation of Bipolar Charge Transport in Graphene by Using a Discontinuous Galerkin Method”. Commun. Comput. Phys., 26 (2019), pp. 114–134.
11. MATLAB, 2021. version 9.10.0 (R2021a), Natick, Massachusetts: The MathWorks Inc.
12. Tomadin, A., Brida, D., Cerullo, G., Ferrari, A. C., Polini, M.,: “Nonequilibrium dynamics of photoexcited electrons in graphene: collinear scattering, Auger processes, and the impact of screening’, Phys. Rev. B, Volume 88, 035430 (2013).

On the Discretization of Diffusion Fluxes for a System of PDEs

Falco Schneider

Abstract We consider a system of two PDEs with diffusion type fluxes and discontinuous coefficients. Extending on the ideas of the simple diffusion problem, we derive a two-point flux approximation for a cell centered finite volume method with minimal stencil on a regular Cartesian grid. The proposed approximation takes the coupled characteristic of the original problem into account and is compared to a commonly deployed decoupled approximation using separate harmonic averages of the transport coefficients. Equivalence of both methods is shown if the coefficients fulfill some linear relation and generalization to non-uniform rectilinear grids is discussed. We conclude our analysis by performing a numerical study for our specific application of liquid electrolytes in Li-ion batteries.

1 Introduction

The cell centered finite volume method for scalar diffusion problems with discontinuous coefficients is commonly deployed using a two-point flux approximation based on the harmonic average of the coefficient. Given two neighbouring cells i, j of size h, see Fig. 1, the flux approximation $\mathbf{N}_{i,j}$ between the two cells is constructed by imposing continuity at the cell interfaces for the concentration c and the discrete diffusion fluxes $\mathbf{N}_i, \mathbf{N}_j$ of the respective cells, while assuming that the diffusion coefficient D is constant within each cell. This leads to the expression

$$\mathbf{N}_{i,j} \cdot \mathbf{n} = \mathbf{N}_i \cdot \mathbf{n} = \mathbf{N}_j \cdot \mathbf{n} = -D_{i,j}\frac{c_j - c_i}{h} = -\frac{2}{D_i^{-1} + D_j^{-1}}\frac{c_j - c_i}{h}, \tag{1}$$

where $D_{i,j}$ is the harmonic average of $D_i, D_j > 0$ and $\mathbf{n}$ the normal from cell i to j.

F. Schneider (✉)
Fraunhofer ITWM, Kaiserslautern, Germany
e-mail: falco.schneider@itwm.fraunhofer.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_38

In this paper we want to consider systems of PDEs of the form

$$\partial_t c = -\nabla \cdot \mathbf{N}(c, \phi), \quad 0 = -\nabla \cdot \mathbf{J}(c, \phi), \quad x \in \Omega \subset \mathbb{R}^3, \ t \in [0, T_{\text{fin}}], \tag{2}$$

where the fluxes **N**, **J** of the two unknown quantities c, ϕ are given by

$$\begin{pmatrix} \mathbf{N} \\ \mathbf{J} \end{pmatrix} = \Lambda \cdot \begin{pmatrix} \nabla c \\ \nabla \phi \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \zeta \end{pmatrix} \cdot \begin{pmatrix} \nabla c \\ \nabla \phi \end{pmatrix} = \begin{pmatrix} \mathbf{N}^c \\ \mathbf{J}^c \end{pmatrix} + \begin{pmatrix} \mathbf{N}^\phi \\ \mathbf{J}^\phi \end{pmatrix}, \tag{3}$$

such that they are coupled by a potentially asymmetric matrix Λ of discontinuous coefficients. This is motivated by our application of Li-ion batteries [2], where the transport of ions and charge in liquid electrolytes are directly coupled, such that the Li-ion flux density **N** as well as the current density **J** depend on the gradient of the Li-ion concentration c and the electrochemical potential ϕ. The coupled flux approximation is not limited to the parabolic-elliptic system (2) and might be applied to similar systems using fluxes of the form (3).

The paper is structured as follows: In Sect. 2 we derive the coupled two-point flux approximation, while Sect. 3 introduces a commonly used decoupled approximation. We discuss equivalence of both methods for a specific setting and generalization to non-uniform rectilinear grids in Sect. 4. Finally, Sect. 5 presents numerical results of an electrolyte simulation, followed by a conclusion in Sect. 6.

2 Coupled Flux Discretization

Motivated by our application, we want to consider general fluxes of the form (3), where we make two assumptions on the coefficient matrix

1. The coefficient matrix Λ is non-singular,
2. The matrix sum $\Lambda_i + \Lambda_j$ for arbitrary cells i, j is non-singular,

which are sufficient for the coupled flux approximation to be well defined. Given two adjacent cells i, j with their respective values for c, ϕ, Λ, we impose continuity of the variables c, ϕ and fluxes **N**, **J** at the interface, while the coefficients are assumed to be constant in each cell, see Fig. 1. The discrete fluxes for each cell read

$$\begin{pmatrix} \mathbf{N_i} \cdot \mathbf{n} \\ \mathbf{J_i} \cdot \mathbf{n} \end{pmatrix} = \frac{2\Lambda_i}{h} \cdot \begin{pmatrix} c_w - c_i \\ \phi_w - \phi_i \end{pmatrix}, \qquad \begin{pmatrix} \mathbf{N_j} \cdot \mathbf{n} \\ \mathbf{J_j} \cdot \mathbf{n} \end{pmatrix} = \frac{2\Lambda_j}{h} \cdot \begin{pmatrix} c_j - c_w \\ \phi_j - \phi_w \end{pmatrix}, \tag{4}$$

and are determined by the unknown interface values c_w, ϕ_w of the continuous variables c, ϕ. From the continuity of the fluxes $\mathbf{N_i} \cdot \mathbf{n} = \mathbf{N_j} \cdot \mathbf{n}$ and $\mathbf{J_i} \cdot \mathbf{n} = \mathbf{J_j} \cdot \mathbf{n}$ we get a linear system $Av = b$ for the interface quantities $v = (c_w, \phi_w)$, with A given by $\Lambda_i + \Lambda_j$ and Assumption 2 guarantees the existence of a unique solution. By solving for c_w, ϕ_w and plugging the obtained values back into the definition of the fluxes (4), we obtain the flux approximations

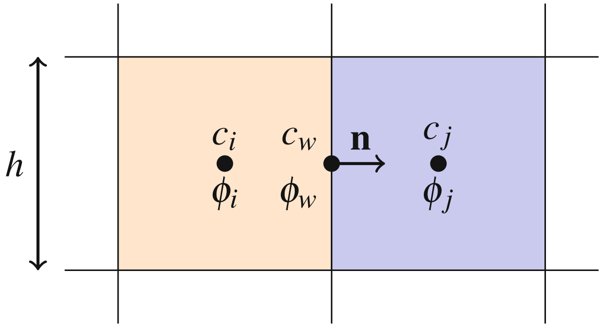


Fig. 1 Cell centered finite volume setup on regular Cartesian grid. The flux at centers of cell interfaces is approximated via the cell centered values of the adjacent cells

$$\begin{pmatrix} \mathbf{N_{i,j}} \cdot \mathbf{n} \\ \mathbf{J_{i,j}} \cdot \mathbf{n} \end{pmatrix} = \begin{pmatrix} \mathbf{N_i} \cdot \mathbf{n} \\ \mathbf{J_i} \cdot \mathbf{n} \end{pmatrix} = \begin{pmatrix} \mathbf{N_j} \cdot \mathbf{n} \\ \mathbf{J_j} \cdot \mathbf{n} \end{pmatrix} = \frac{\Lambda_{i,j}}{h} \cdot \begin{pmatrix} c_j - c_i \\ \phi_j - \phi_i \end{pmatrix}, \tag{5}$$

where the coefficient matrix is approximated by the matrix harmonic average

$$\begin{pmatrix} \alpha_{i,j} & \beta_{i,j} \\ \gamma_{i,j} & \zeta_{i,j} \end{pmatrix} = \Lambda_{i,j} = 2\left(\Lambda_i^{-1} + \Lambda_j^{-1}\right)^{-1}. \tag{6}$$

Thus, the coupled approach yields a flux approximation with a structure very similar to the scalar diffusion problem. This matches other results reported in the literature, e.g. [4], where Λ is assumed to be symmetric positive definite.

3 Decoupled Flux Discretization

Based on the result (1), one might try to approximate the flux of the coupled problem (3) by using the harmonic average for each of the four coefficients separately

$$\alpha_{i,j} = \frac{2}{\alpha_i^{-1} + \alpha_j^{-1}}, \ \beta_{i,j} = \frac{2}{\beta_i^{-1} + \beta_j^{-1}}, \ \gamma_{i,j} = \frac{2}{\gamma_i^{-1} + \gamma_j^{-1}}, \ \zeta_{i,j} = \frac{2}{\zeta_i^{-1} + \zeta_j^{-1}}, \tag{7}$$

and plug them into (5). This is generally different to the coupled approach (6). Each coefficient approximation in the decoupled approach only depends on the respective transport coefficient, while in the coupled approach one generally obtains a dependency on all four transport coefficients. This approach corresponds to imposing continuity for the fluxes **N** and **J**, but also enforcing continuity for each gradient term in the fluxes separately. Using the notation from (3), the constraints read

$$\mathbf{N}_i^c \cdot \mathbf{n} = \mathbf{N}_j^c \cdot \mathbf{n}, \quad \mathbf{J}_i^c \cdot \mathbf{n} = \mathbf{J}_j^c \cdot \mathbf{n}, \quad \mathbf{N}_i^\phi \cdot \mathbf{n} = \mathbf{N}_j^\phi \cdot \mathbf{n}, \quad \mathbf{J}_i^\phi \cdot \mathbf{n} = \mathbf{J}_j^\phi \cdot \mathbf{n}. \tag{8}$$

Since we have a total of four equations, we assume separate interface quantities c_{w_1}, ϕ_{w_1} and c_{w_2}, ϕ_{w_2} for **N** and **J**, respectively. The equations (8) can be solved independently and analogously to the scalar diffusion problem. Thus, we obtain the averages of the coefficients as in (7) and the interface quantities

$$c_{w_1} = \frac{\alpha_i c_i + \alpha_j c_j}{\alpha_i + \alpha_j}, \; c_{w_2} = \frac{\gamma_i c_i + \gamma_j c_j}{\gamma_i + \gamma_j}, \; \phi_{w_1} = \frac{\beta_i \phi_i + \beta_j \phi_j}{\beta_i + \beta_j}, \; \phi_{w_2} = \frac{\zeta_i \phi_i + \zeta_j \phi_j}{\zeta_i + \zeta_j}, \tag{9}$$

which will generally be inconsistent between **N** and **J** for the decoupled approach.

4 Equivalence of Both Approaches and Rectilinear Grids

In general, we have $c_{w_1} \neq c_{w_2}$, $\phi_{w_1} \neq \phi_{w_2}$, for arbitrary values of c_i, c_j, ϕ_i, ϕ_j and the coupled approach will be different to the decoupled approach, using separate harmonic averages. However, it is easy to see that these interface quantities will be consistent if we have some linear relation between the coefficients.

If $\alpha = k_1 \gamma$ for some $k_1 \in \mathbb{R} \setminus \{0\}$, we obtain $c_{w_1} = c_{w_2}$. Analogously, $\beta = k_2 \zeta$ for some $k_2 \in \mathbb{R} \setminus \{0\}$ implies $\phi_{w_1} = \phi_{w_2}$. If both of these relations are fulfilled, we know by Assumption 1 that $k_1 \neq k_2$ and can even show equivalence to the interface quantities of the coupled approach

$$c_{w_1} = c_{w_2} = c_w, \qquad \phi_{w_1} = \phi_{w_2} = \phi_w, \tag{10}$$

by simplifying the explicit expressions obtained for c_w, ϕ_w. In particular, we obtain that (4), (5) and (6) hold for the decoupled approach and both methods coincide.

The obtained results can be generalized to non-uniform rectilinear grids in a straightforward manner. In that case, both flux approximations are again structurally similar and of the form

$$\begin{pmatrix} \mathbf{N}_{i,j} \cdot \mathbf{n} \\ \mathbf{J}_{i,j} \cdot \mathbf{n} \end{pmatrix} = \begin{pmatrix} \alpha_{i,j} & \beta_{i,j} \\ \gamma_{i,j} & \zeta_{i,j} \end{pmatrix} \cdot \begin{pmatrix} \frac{c_j - c_i}{d_{i,j}} \\ \frac{\phi_j - \phi_i}{d_{i,j}} \end{pmatrix}, \tag{11}$$

with the distance between cell centers $d_{i,j} = 0.5(h_i + h_j)$. For the decoupled approach, one obtains the weighted harmonic average of each individual coefficient with the weights given by the respective cell sizes. The coefficient approximation of the coupled approach is given by the weighted matrix harmonic average

$$\Lambda_{i,j} = (h_i + h_j) \left(h_i \Lambda_i^{-1} + h_j \Lambda_j^{-1} \right)^{-1}, \tag{12}$$

where we also need to modify Assumption 2 to ensure the matrix sum $h_i^{-1}\Lambda_i + h_j^{-1}\Lambda_j$ to be non-singular for arbitrary cells i, j, such that the approach is well defined. Note, equivalence of the two approximations for the aforementioned linear relations still holds on these type of grids.

5 A Numerical Study for Liquid Electrolytes

We conduct a spatial convergence analysis to compare the two approximations (6) and (7) using a one-dimensional electrolyte solver. The corresponding system reads

$$\partial_t c = -\partial_x N(c,\phi), \quad 0 = -\partial_x J(c,\phi), \quad x \in [0,L],\ t \in [0,T_{\text{fin}}], \tag{13}$$

with constant initial concentration profile $c_0 \equiv 1\,\text{mol/l}$. We apply constant Neumann boundary conditions $J|_{x=0} = J|_{x=L} = 45\,\text{mA/cm}^2$ to simulate a steady current flowing through the domain and a consistent ion flux $N|_{x=0} = N|_{x=L} = F^{-1}J|_{x=0}$. For the potential to be uniquely defined and centered around 1 V, we additionally prescribe a mean condition. The isothermal coefficients of the battery model [2] are

$$\begin{aligned} \alpha(c) &= -D(c) - RT\tfrac{t_+(c)(t_+(c)-1)}{F^2\,c}\kappa(c), & \beta(c) &= -\tfrac{t_+(c)}{F}\kappa(c),\\ \gamma(c) &= -RT\tfrac{t_+(c)-1}{F\,c}\kappa(c), & \zeta(c) &= -\kappa(c), \end{aligned} \tag{14}$$

where $T = 296\,\text{K}$ is the temperature, R is the gas constant and F is the Faraday constant. Introducing the reference scales $c_{\text{ref}} = 1\,\text{mol/l}$, $\phi_{\text{ref}} = 1\,\text{V}$ and the non-dimensionalized concentration $\hat{c} = c/c_{\text{ref}}$, we use the conductivity from [1]

$$\kappa(c) = \kappa(\hat{c}(c)) = 0.1\frac{\text{S}}{\text{m}} \cdot \left(2.667\cdot\hat{c}^3 - 12.983\cdot\hat{c}^2 + 17.919\cdot\hat{c} + 1.726\right). \tag{15}$$

For the dimensionless transference number we prescribe the function

$$t_+(c) = t_+(\hat{c}(c)) = -0.19333333\cdot\hat{c}^3 + 0.67\cdot\hat{c}^2 - 0.79666667\cdot\hat{c} + 0.58, \tag{16}$$

which is an approximation of the data plotted in Figure 5.3 from [3]. The diffusion coefficient is given by the Einstein relation [1]

$$D(c) = \frac{RT}{F^2\,c}\kappa(c). \tag{17}$$

For physical feasible parameters $\kappa > 0$, $t_+ \in (0,1)$ and D given by the relation (17), we have $\alpha, \beta, \zeta < 0$ and $\gamma > 0$. This guarantees that both assumptions of our coupled approach are fulfilled, because $\det(\Lambda) > 0$ and $\det(\Lambda_i + \Lambda_j) > 0$.

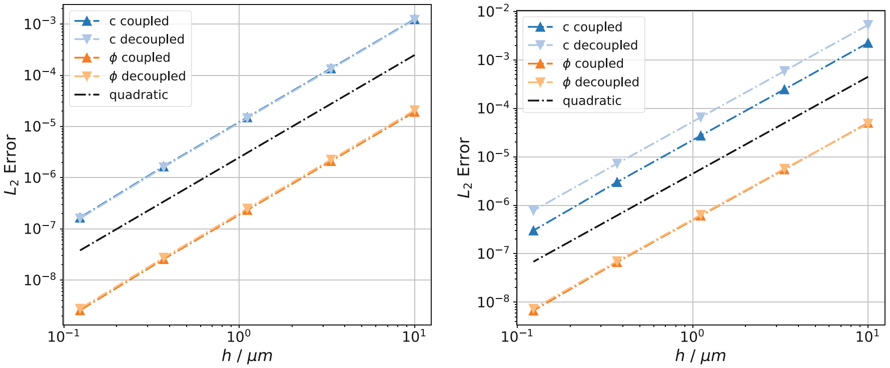


Fig. 2 Spatial convergence of coupled and decoupled flux approximation for 1D electrolyte simulation with respect to the cell size h using a t_+ with low variation (left) or high variation (right)

All simulations use an implicit Euler time stepper with final time $T_{\text{fin}} = 10\,\text{s}$, time step size $\Delta t = 0.1\,\text{s}$ and a domain size of $L = 100\,\mu\text{m}$. We consider uniform discretizations with $N_x \in \{10, 30, 90, 270, 810\}$ cells and compare them to a reference solution with $N_x = 2430$ cells. The nonlinear algebraic systems are solved with a Newton-Raphson method applying a direct linear solver. For the error we consider the following L_2-norm over all common nodes of the space-time grid

$$||u - \tilde{u}||_{L_2} = \frac{1}{u_{\text{ref}}} \left(\frac{1}{N_t N_x} \sum_{i,j=1}^{N_x, N_t} |u_{i,j} - \tilde{u}_{i,j}|^2 \right)^{0.5}, \qquad (18)$$

where the error is calculated separately for c and ϕ and normalized by the reference scales c_{ref}, ϕ_{ref}, respectively. The resulting convergence plot can be seen on the left in Fig. 2. It turns out, if we assume t_+ to be constant, then the coefficients of the battery model fulfill the aforementioned linear relations, due to the Einstein relation between D and κ. In this case, both approaches are equivalent. Considering (16), t_+ shows only small variations, such that the errors of the methods are still similar. If we switch to an expression $t_+(\hat{c}(c)) = 0.5 - 0.4\tanh(6(\hat{c} - 1))$ with stronger variations, we obtain the plot on the right in Fig. 2, where the coupled method is more accurate. Overall, we observe second order convergence for both methods.

6 Conclusion

We derived a coupled flux approximation for systems of two coupled gradient fluxes, which takes the coupled structure of the fluxes into account and is generally different to a decoupled approach using separate harmonic averages. Structurally, the two approaches are similar, even equivalent if there exists a specific linear relation between the coefficients. For our example of an electrolyte, the two methods coincide for t_+ constant. With growing variations in $t_+(c)$, larger differences between the methods are observed, where the coupled approach is generally more accurate. Thus, the method might be advantageous for other problems where the coefficients do not match the aforementioned linear relations.

Acknowledgments I would like to thank Dr. Heiko Andrä, Prof. Oleg Iliev, Prof. Axel Klar and Dr. Jochen Zausch for the interesting scientific discussions on this topic, as well as the valuable suggestions of scientific and editorial nature.

References

1. Ecker, M., Tran, T.K.D., Dechent, P., Käbitz, S., Warnecke, A., Sauer, D.U.: Parameterization of a Physico-Chemical Model of a Lithium-Ion Battery: I. Determination of Parameters. Journal of The Electrochemical Society **162**(9), A1836–A1848 (2015).
2. Latz, A., Zausch, J.: Thermodynamic consistent transport theory of Li-ion batteries. Journal of Power Sources **196**(6), 3296–3302 (2011).
3. Nyman, A.: An Experimental and Theoretical Study of the Mass Transport in Lithium-Ion Battery Electrolytes. Ph.D. thesis, KTH, Applied Electrochemistry (2011)
4. Thije Boonkkamp, ten, J., Liu, L., Dijk, van, J., Peerenboom, K.: Harmonic complete flux schemes for conservation laws with discontinuous coefficients, CASA-report, vol. 1329. Technische Universiteit Eindhoven (2013)

Dynamics of the N-fold Pendulum in the Framework of Lie Group Integrators

Elena Celledoni, Ergys Çokaj, Andrea Leone, Davide Murari,
and Brynjulf Owren

Abstract Since their introduction, Lie group integrators have become a method of choice in many application areas. Various formulations of these integrators exist, and in this work we focus on Runge-Kutta-Munthe-Kaas methods. First, we briefly introduce this class of integrators, considering some of the practical aspects of their implementation, such as adaptive time stepping. We then present some mathematical background that allows us to apply them to some families of Lagrangian mechanical systems. We conclude with an application to a nontrivial mechanical system: the N-fold 3D pendulum.

1 Introduction

Lie group integrators are used to simulate problems whose solution evolves on a manifold. Many approaches to Lie group integrators can be found in the literature, with several applications for mechanical systems (see, e.g. [2, 8, 9]).

The present work is motivated by applications in modelling and simulation of slender structures like beams, and the example considered here is a chain of pendulums. The dynamics of this mechanical system is described in terms of a Lie group G acting transitively on the phase space $\mathcal{M}$. This setting is used to build also a numerical integrator.

In Sect. 2 we give a brief overview of the Runge-Kutta-Munthe-Kaas (RKMK) methods with particular focus on the variable step size methods, which we use later in Sect. 4.2 for the numerical experiments. In Sect. 3 we introduce some necessary mathematical background that allows us to apply RKMK methods to the system of interest. In particular, we focus on a condition that guarantees the homogeneity of

E. Celledoni · E. Çokaj (✉) · A. Leone · D. Murari · B. Owren
Department of Mathematical Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway
e-mail: elena.celledoni@ntnu.no; ergys.cokaj@ntnu.no; andrea.leone@ntnu.no; davide.murari@ntnu.no; brynjulf.owren@ntnu.no

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_39

the tangent bundle TQ of a manifold Q. We then consider Cartesian products of homogeneous manifolds. In Sect. 4 we reframe the ODE system of the chain of N connected $3D$ pendulums in the geometric framework presented in Sect. 3. We write the equations of motion and represent them in terms of the infinitesimal generator of the transitive action. The final part shows some numerical experiments where the constant and variable step size methods are compared.

2 RKMK Methods with Variable Step Size

The underlying idea of RKMK methods is to express a vector field $F \in \mathfrak{X}(\mathcal{M})$ as $F|_m = \psi_*(f(m))|_m$, where ψ_* is the infinitesimal generator of ψ, a transitive action on $\mathcal{M}$, and $f\colon \mathcal{M} \to \mathfrak{g}$. This allows us to transform the problem from the manifold $\mathcal{M}$ to the Lie algebra $\mathfrak{g}$, on which we can perform a time step integration. We then map the result back to $\mathcal{M}$, and repeat this up to the final integration time. More explicitly, let h_n be the size of the $n-$th time step, we then update $y_n \in \mathcal{M}$ to y_{n+1} by

$$\begin{cases} \sigma(0) = 0 \in \mathfrak{g}, \\ \dot{\sigma}(t) = \mathrm{dexp}^{-1}_{\sigma(t)} \circ f \circ \psi(\exp(\sigma(t)), y_n) \in T_{\sigma(t)}\mathfrak{g}, \\ y_{n+1} = \psi(\exp(\sigma_1), y_n) \in \mathcal{M}, \end{cases} \tag{1}$$

where $\sigma_1 \approx \sigma(h_n) \in \mathfrak{g}$ is computed with a Runge-Kutta method.

One approach for varying the step size is based on embedded Runge-Kutta pairs for vector spaces. This approach consists of a principal method of order p, used to propagate the numerical solution, together with some auxiliary method, of order $\tilde{p} < p$, that is only used to obtain an estimate of the local error. This local error estimate is in turn used to derive a step size adjustment formula that attempts to keep the local error estimate approximately equal to some user-defined tolerance tol in every step. Both methods are applied to solve the ODE for $\sigma(t)$ in (1), yielding two approximations σ_1 and $\tilde{\sigma}_1$ respectively, using the same step size h_n. Now, some distance measure between σ_1 and $\tilde{\sigma}_1$ provides an estimate e_{n+1} for the size of the local truncation error. Thus, $e_{n+1} = Ch_{n+1}^{\tilde{p}+1} + \mathcal{O}(h^{\tilde{p}+2})$. Aiming at $e_{n+1} \approx$ tol in every step, one may use a formula of the type

$$h_{n+1} = \theta \left(\frac{\mathrm{tol}}{e_{n+1}}\right)^{\frac{1}{\tilde{p}+1}} h_n, \tag{2}$$

where θ is typically chosen between 0.8 and 0.9. If $e_n >$ tol, the step is rejected. Hence, we can redo the step with the step size obtained by the same formula.

3 Mathematical Background

This section introduces the mathematical background that allows us to study many mechanical systems in the framework of Lie group integrators and Lie group actions. In particular, we provide some results that we use to study the model of a chain of N 3D-pendulums presented in the last section.

3.1 The Tangent Bundle of Some Homogeneous Manifolds Is Homogeneous

For Lagrangian mechanical systems, the phase space is usually the tangent bundle TQ of some configuration manifold Q. In [1] the authors present a setting in which the homogeneity of Q implies that of TQ. We now briefly review and reframe it in the notation used throughout the paper.

Consider a smooth homogeneous $n-$dimensional manifold Q. This means that Q is endowed with a transitive G-group action $\Lambda\colon G \times Q \to Q$, i.e., for any pair $q_1, q_2 \in Q$ there is $g \in G$ such that $\Lambda(g, q_1) = q_2$. Assume that for each $q \in Q$, the map $\Lambda_q : G \to Q$ defined as $\Lambda_q(g) := \Lambda(g, q)$, is a submersion at $e \in G$. When these hypotheses hold, it can be shown that TQ is a homogeneous manifold as well, and an explicit transitive action can be obtained from Λ. Let Λ_* be the infinitesimal generator of the group action Λ, and denote with $\bar{\xi}(q) := \Lambda_*(\xi)(q) \in T_qQ$ the differential at the identity element $e \in G$ of Λ_q, evaluated at $\xi \in \mathfrak{g}$. We then introduce $\Lambda_g\colon Q \to Q$, $q \mapsto \Lambda(g, q)$ and call $T_{\bar{q}}\Lambda_g$ its tangent lift at $\bar{q} \in Q$.

Consider the manifold $\bar{G} := G \ltimes \mathfrak{g}$, equipped with the semi-direct product Lie group structure (see, e.g., [4]). We can introduce a transitive group action on TQ as follows:

$$\varphi\colon \bar{G} \times TQ \to TQ, \quad \big((g, \xi), (q, v)\big) \mapsto \big(\Lambda(g, q), \bar{\xi}(\Lambda(g, q)) + T_q\Lambda_g(v)\big).$$

By direct computation and basic properties of Lie groups (see, e.g., [5]), it can be seen that the action φ is well defined. Since the action Λ is transitive on Q and Λ_q is assumed to be a submersion at $e \in G$, we have that

$$\forall v' \in T_{q'}Q \;\; \exists \xi \in \mathfrak{g} \text{ s.t. } \Lambda_*(\xi)(q') = \bar{\xi}(\Lambda(g, q)) = v' - T_q\Lambda_g(v).$$

Thus, we conclude that $\mathcal{M} = TQ$ is a homogeneous manifold.

In the application treated in the next section, we are interested in the case in which $Q = S^2 \subset \mathbb{R}^3$, i.e., the unit sphere. In this setting, a transitive group action Λ is given by

$$\Lambda\colon SO(3) \times S^2 \to S^2, \;\; (R, q) \mapsto Rq,$$

$$T_qS^2 \ni \Lambda_*(\xi)(q) = \bar{\xi}(q) = \xi \times q, \quad T_q\Lambda_R(v) = Rv \in T_{Rq}S^2.$$

Therefore, in this case we recover the restriction to $TS^2 \subset \mathbb{R}^6 \simeq \mathfrak{se}(3)$ of the Adjoint action of $\bar{G} = SE(3) = SO(3) \ltimes \mathbb{R}^3 \simeq SO(3) \ltimes \mathfrak{so}(3)$ (see, e.g., [6])

$$\varphi((R,r),(q,v)) = (Rq, Rv + r \times Rq) = {}^{1}(Rq, Rv + \hat{r}Rq), \tag{3}$$

which hence becomes a particular case of a more general framework.

3.2 The Cartesian Product of Homogeneous Manifolds Is Homogeneous

Consider a family of homogeneous manifolds $\mathcal{M}_1, \dots, \mathcal{M}_n$. Call $(G_i, \odot_i)$ the Lie group acting transitively on the associated smooth manifold $\mathcal{M}_i$, and φ_i such a transitive action. Let $\mathfrak{g}_i$ be the Lie algebra of G_i, $i = 1, \dots, n$, and

$$\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_n, \quad G = G_1 \times G_2 \times \cdots \times G_n.$$

The manifold G can be naturally equipped with a Lie group structure given by the direct product. More precisely, for a pair of elements $G \ni g_i = (g_i^1, \dots, g_i^n)$, $i = 1, 2$, we can define their product $g_1 \cdot g_2 := (g_1^1 \odot_1 g_2^1, \dots, g_1^n \odot_n g_2^n) \in G$. We can similarly define componentwise the exponential map.

This construction ensures that the manifold $\mathcal{M}$ is homogeneous too, and G acts transitively on it. That is, let

$$g = (g^1, \dots, g^n) \in G, \quad m = (m^1, \dots, m^n) \in \mathcal{M},$$

then

$$\varphi\colon G \times \mathcal{M} \to \mathcal{M}, \quad \varphi(g, m) := (\varphi_1(g^1, m^1), \dots, \varphi_n(g^n, m^n)).$$

We now restrict to the specific case $\mathcal{M}_i = TS^2$ for $i = 1, \dots, n$. Since TS^2 is a homogeneous manifold with transitive action φ defined as in Eq. (3), we can write the transitive group action

$$\psi\colon (SE(3))^n \times (TS^2)^n \to (TS^2)^n,$$

$$\psi\big((g^1, \dots, g^n), (m^1, \dots, m^n)\big) = \big(\varphi(g^1, m^1), \dots, \varphi(g^n, m^n)\big),$$

where $g^i := (R_i, r_i) \in SE(3)$, $m^i = (q_i, v_i) \in TS^2$.

[1] Here $\hat{r} = \begin{bmatrix} 0 & -r_3 & r_2 \\ r_3 & 0 & -r_1 \\ -r_2 & r_1 & 0 \end{bmatrix}$, where $r = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}$.

4 The N-fold 3D Pendulum

We now apply the geometric setting from Sect. 3 to the specific problem of a chain of N connected 3D pendulums, whose dynamics evolves on $(TS^2)^N$.

4.1 Equations of Motion

Let us consider a chain of N pendulums subject to constant gravity g. The system is modeled by N rigid, massless links serially connected by spherical joints, with the first link connected to a fixed point placed at the origin of the ambient space $\mathbb{R}^3$, as in Fig. 1. We neglect friction and interactions among the pendulums.

The modeling part comes from [7] and we omit details. We denote by $q_i \in S^2$ the configuration vector of the $i-$th mass, m_i, of the chain. Following [7], we express the Euler–Lagrange equations for our system in terms of the configuration variables $(q_1, \dots, q_N) \in (S^2)^N \subset \mathbb{R}^{3N}$, and their angular velocities $(\omega_1, \dots, \omega_N) \in T_{q_1}S^2 \times \cdots \times T_{q_N}S^2 \subset \mathbb{R}^{3N}$, defined be the following kinematic equations:

$$\dot{q}_i = \omega_i \times q_i, \quad i = 1, \dots, N. \tag{4}$$

The Euler–Lagrange equations of the system can be written as

$$R(q)\dot{\omega} = \begin{bmatrix} \sum\limits_{\substack{j=1\\ j\neq i}}^{N} M_{ij}|\omega_j|^2\hat{q}_i q_j - \Big(\sum\limits_{j=i}^{N} m_j\Big) g L_i \hat{q}_i e_3 \end{bmatrix}_{i=1,\dots,N} = \begin{bmatrix} r_1 \\ \vdots \\ r_N \end{bmatrix} \in \mathbb{R}^{3N}, \tag{5}$$

where $R(q) \in \mathbb{R}^{3N\times 3N}$ is a symmetric block matrix defined as

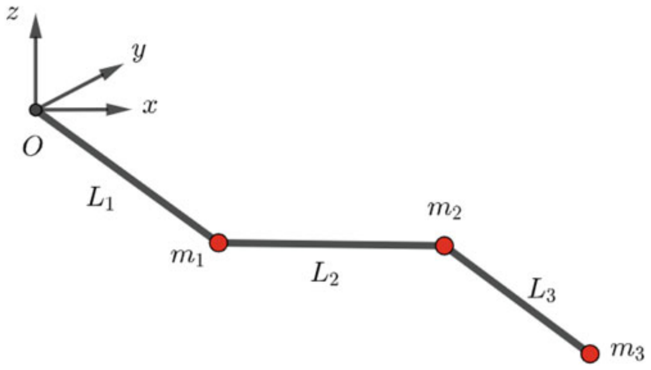


Fig. 1 Chain of 3 connected pendulums at a fixed time instant

$$R(q)_{ii} = \Big(\sum_{j=i}^{N} m_j\Big) L_i^2 I_3 \in \mathbb{R}^{3\times 3},$$

$$R(q)_{ij} = \Big(\sum_{k=j}^{N} m_k\Big) L_i L_j \hat{q}_i^T \hat{q}_j \in \mathbb{R}^{3\times 3} = R(q)_{ji}^T,\ i < j,$$

and

$$M_{ij} = \Big(\sum_{k=\max\{i,j\}}^{N} m_k\Big) L_i L_j I_3 \in \mathbb{R}^{3\times 3}.$$

Equations (4) and (5) define the dynamics of the N-fold pendulum, and hence a vector field $F \in \mathfrak{X}((TS^2)^N)$. We now find a function $f\colon (TS^2)^N \to \mathfrak{se}(3)^N$ such that

$$\psi_*(f(m))|_m = F|_m, \quad \forall m \in (TS^2)^N,$$

where ψ is defined as in Sect. 3.2.

Since $R(q)$ defines a linear invertible map (see [2])

$$A_q\colon T_{q_1}S^2 \times \cdots \times T_{q_N}S^2 \to T_{q_1}S^2 \times \cdots \times T_{q_N}S^2, \quad A_q(\omega) := R(q)\omega,$$

we can rewrite the ODEs for the angular velocities as follows:

$$\dot{\omega} = A_q^{-1}\left(\begin{bmatrix} r_1 \\ \vdots \\ r_N \end{bmatrix}\right) = \begin{bmatrix} h_1(q,\omega) \\ \vdots \\ h_N(q,\omega) \end{bmatrix} = \begin{bmatrix} a_1(q,\omega) \times q_1 \\ \vdots \\ a_N(q,\omega) \times q_N \end{bmatrix}. \tag{6}$$

In equation (6) the r_is are defined as in (5), and $a_1, \ldots, a_N\colon (TS^2)^N \to \mathbb{R}^3$ can be defined as $a_i(q,\omega) := q_i \times h_i(q,\omega)$. Thus, the map f is given by

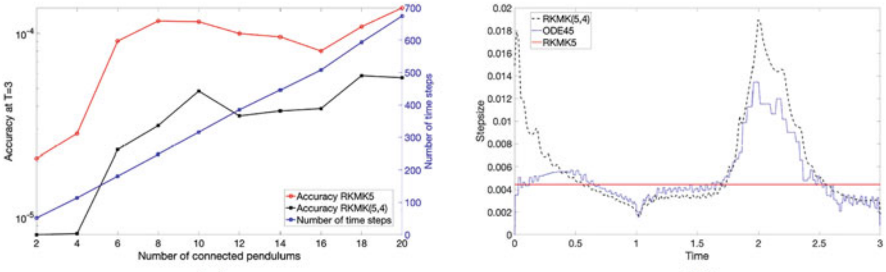
$$f(q,\omega) = \begin{bmatrix} \omega_1 \\ q_1 \times h_1(q,\omega) \\ \vdots \\ \omega_N \\ q_N \times h_N(q,\omega) \end{bmatrix} \in \mathfrak{se}(3)^N \simeq \mathbb{R}^{6N}.$$

4.2 Numerical Experiments

In this section we show a numerical experiment with the N-fold 3D pendulum, in which we compare the performance of constant and variable step size methods. We do not show results on the preservation of the geometry (up to machine accuracy), since this is given by construction. We consider the RKMK pair coming from Dormand–Prince method (DOPRI 5(4) [3], which we denote by RKMK(5,4)). We set a tolerance of 10^{-6} and solve the system with the RKMK(5,4) scheme. Fixing the number of time steps required by RKMK(5,4), we repeat the experiment with RKMK of order 5 (denoted by RKMK5). The comparison occurs at the final time $T = 3$ using the Euclidean norm of the ambient space $\mathbb{R}^{6N}$. The quality of the approximation is measured against a reference solution obtained with ODE45 from MATLAB with a strict tolerance.

The motivating application behind the choice of this mechanical system has been some intuitive relation with flexible slender structures like beams. For this limiting behaviour to make sense, we first fix the length of the entire chain of pendulums to some L, then we set the size of each pendulum to $L_i = L/N$ and initialize $(q_i, \omega_i) = (1, 0, 0, 0, 0, 0)$, $\forall i = 1, \ldots, N$. As we can see in Fig. 2a, the results of our experiments show that number of time steps that RKMK(5,4) requires to reach the desired accuracy increases with N, and this can be read in terms of an augmentation of the dynamics' complexity. For this reason, as highlighted in Fig. 2, distributing these time steps uniformly in the time interval $[0, T]$ becomes an inefficient approach, and hence a variable step size method gives better performance.

We further design a slightly different experiment to compare the computational time of the constant and variable step size RKMK methods. First, we fix the tolerance $tol = 10^{-6}$ for RKMK(5,4) and compute its distance from the reference solution with ODE45. Then, we aim to replicate this error with RKMK5, increasing the number of performed time steps. We report in Table 1 the results of the experiment. Because of the more efficient distribution of the time steps, we notice smaller values with RKMK(5,4) for the more involved systems.



(a)

(b)

Fig. 2 Comparisons of variable versus constant step size for the N-fold 3D pendulum. (**a**) Accuracy against the number of pendulums. (**b**) Comparison of step sizes with 20 pendulums

Table 1 Elapsed times (in seconds) obtained with RKMK5 (second row) and with RKMK(5,4) (third row) for systems having different number of pendulums (first row). In the last row we report the ratio between the RKMK5 and the RKMK(5,4) runtimes. These are obtained with the `tic–toc` command of MATLAB

Pendulums	2	4	6	8	10	12	14	16	18	20
RKMK5	0.12	0.42	1.04	2.24	3.80	6.74	9.09	12.71	18.51	27.67
RKMK(5,4)	0.16	0.38	0.91	1.59	2.83	4.51	6.93	9.71	13.68	18.81
Ratio	0.75	1.11	1.14	1.41	1.34	1.49	1.31	1.31	1.35	1.47

Acknowledgments This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860124.

References

1. Brockett, R. W., Sussmann, H. J. (1972). Tangent bundles of homogeneous spaces are homogeneous spaces. In Proc. Amer. Math. Soc., 35(2),550–551.
2. Celledoni, E., Çokaj, E., Leone, A., Murari, D., Owren, B. (2021). Lie Group integrators for mechanical systems. International Journal of Computer Mathematics.
3. Dormand, J. R., Prince, P. J. (1980). A family of embedded Runge-Kutta formulae. Journal of Computational and Applied Mathematics, 6(1), 19–26.
4. Engø, K. (2003). Partitioned Runge–Kutta methods in Lie-group setting. BIT Numerical Mathematics, 43(1), 21–39.
5. Hall, B. (2015). Lie groups, Lie algebras, and representations: an elementary introduction. Springer, 222.
6. Holm, D. D., Schmah, T., Stoica, C. (2009). Geometric mechanics and symmetry: from finite to infinite dimensions, 12. Oxford University Press.
7. Lee, T., Leok, M., McClamroch, N. H. (2018). Global formulations of Lagrangian and Hamiltonian dynamics on manifolds. Interaction of Mechanics and Mathematics. Springer, Cham. A geometric approach to modeling and analysis.
8. Iserles, A., Munthe-Kaas, H. Z., Nørsett, S. P., and Zanna, A. (2000). Lie-group methods. Acta Numerica, 9:215–365.
9. Celledoni, E., Marthinsen, H., and Owren, B. (2014). An introduction to Lie group integrators: basics, new developments and applications. J. Comput. Phys., 257(part B):1040–1061.

Hydrodynamic Interaction Between a Row of Oblate Spheroids in a Steady Stream of Viscous Fluid

Tsvetan Kotsev

Abstract This paper presents a numerical simulation of flow structure past three oblate spheroids arranged in line placed in a stream of viscous fluid with uniform velocity parallel to the line connecting the body centers. The hydrodynamic interaction was studied for Reynolds numbers from 1 to 100. The axis ratio of the bodies takes values 0.1, 0.2, 0.5 and 0.8 and the distances between them varies from very small one up to 20 big particle diameters. At some values of Re and distance between spheroids the drag of the middle and downstream ones takes negative values that means a “attractive force” exists between the bodies.

1 Introduction

In many engineering applications, chemical processes, energy and spray systems, etc., the hydrodynamic interaction between solid particles or droplets immersed in moving viscous fluid is quite important. There are lot of studies done both by analytical and numerical methods concerning the viscous flow past one or two particles but most of them treat particles with exactly spherical shape [1–10]. The evolution of the flow field and drag experienced by three spherical particles arranged in line when the distance between them varies was studied numerically in [11] for Reynolds numbers up to 200. Relatively less in the literature are the studies concerning flow past particles with shape different from spherical. Some of them are [12–17]. The goal of the present study is to simulate the flow structure and hydrodynamic interaction between three oblate spheroids in line at Reynolds numbers up to 100.

T. Kotsev (✉)
Sofia University, Department of Mathematics and Informatics, Sofia, Bulgaria
e-mail: tkotsev@fmi.uni-sofia.bg

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_40

2 Formulation of the Problem and Method of Solution

Three rigid spheroidal particles arranged in line are immersed in a stream of incompressible viscous fluid with velocity V_0 at infinity parallel to their central line. The aspect ratio is defined as the ratio between the revolution axis and the major axis $\lambda = c/a$. Particles of equal size and aspect ratio are considered. The numerical simulations were done in Cartesian coordinate system where the centers of the spheroids are located on y-axis (Fig. 1). The fluid is water with density of 998 kg/m^3 and dynamic viscosity 0.00101 Pa.s. The Reynolds number is based on the big diameter of the upstream particle—$\mathrm{Re} = 2a_1.V_0/\nu$, where ν is the kinematic viscosity of the fluid. It is assumed that the flow field behind the particles stays axisymmetric and stable for Reynolds numbers $\mathrm{Re} < 100$ and axis ratio between 0.1 and 1. The distance between upstream and middle spheroid is denoted with d_1 and between the middle and downstream spheroid with d_2. They are non-dimensionalized with respect to the large semi-axis of the particles.

The equations describing the fluid motion are continuity and steady Navier-Stokes equations as follows:

$$\nabla \cdot \mathbf{V} = 0 \tag{1}$$

$$\rho(\mathbf{V} \cdot \nabla)\mathbf{V} = -\nabla p + \mu \nabla^2 \mathbf{V} \ , \tag{2}$$

where $\mathbf{V}$ is velocity vector, μ—dynamic viscosity, p—pressure, ρ—fluid density.

Boundary conditions are:

$$\text{no slip on the spheroid surfaces} - V_x = 0, V_y = 0, V_z = 0, \tag{3}$$

$$\text{uniform velocity at the inlet} - V_x = 0, V_z = 0, V_y = V_0, \tag{4}$$

$$\text{zero pressure at the outlet} - p = 0, \tag{5}$$

$$\text{symmetry at the outer boundaries} - \mathbf{V} \cdot \mathbf{n} = 0 \ . \tag{6}$$

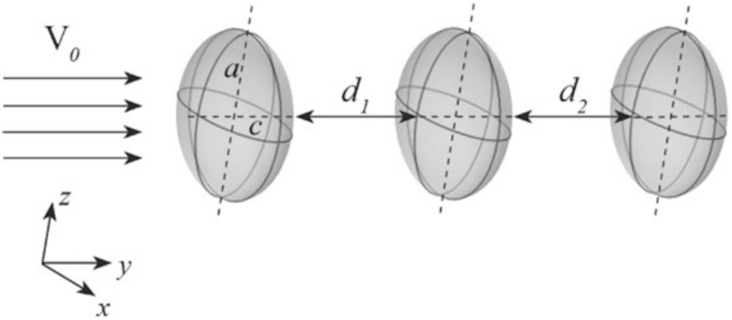


Fig. 1 Geometrical configuration

The drag force is calculated integrating the stress tensor on the body surfaces:

$$F_D = \int_S -p\,\mathbf{n}\cdot\mathbf{j}\,dS + \int_S \mathbf{n}\boldsymbol{\tau}\cdot\mathbf{j}\,dS\,, \tag{7}$$

where S denotes the surface of the spheroids, $\mathbf{j}$ is the unit vector in y direction, $\mathbf{n}$ is the outward normal vector at the surface and $\boldsymbol{\tau}$ is the viscous stress tensor.

The non-dimensional drag coefficient C_D is defined as:

$$C_D = \frac{F_D}{\rho V_0^2 A_{pr}/2}\,, \tag{8}$$

where A_{pr} is the area of the spheroid projected on a plane perpendicular to the flow direction.

Equations (1) and (2) under boundary conditions (3)–(6) are solved numerically using COMSOL Multiphysics package based on the finite element method. The computational domain is 2D due to axial symmetry and is covered with a mesh of triangle elements that approximate in the best way the complicated fluid domain. The outer boundary is set to be at a distance 20 big body diameters far from the line of flow symmetry. Experiments show that this distance is sufficient and moving it further does not give any significance difference in the drag coefficients obtained.

3 Validation of the Computational Scheme

Comparison of the present results with these of Quchene [17] and Happel and Brenner [12] for the flow past a single oblate spheroid at Re $=$ 0.1 is shown in Table 1a). The difference does not exceed 4% and for spheroids with $\lambda = 0.5$ the results are much closer—within 1%. In the case of flow past spheroids being far each from another the present results are also in a good accordance with the results of Juncu [18] for a single spheroid with $\lambda = 0.2$ and 0.5 and Re $=$ 50, 100. The drag values are shown in Table 1b).

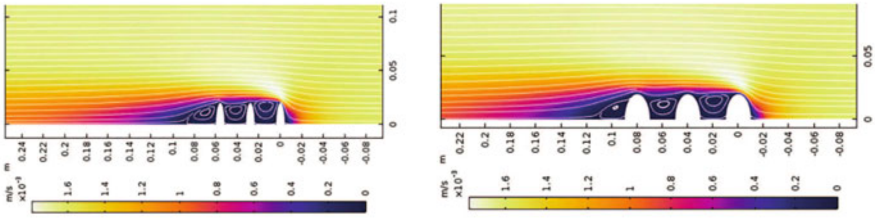
Table 1 Drag coefficients of a single spheroid

λ	Happel and Brenner [12]	Ouchene [17]	Present		Re = 50		Re = 100	
0.2	270.690	281.75	271.88	λ	0.2	0.5	0.2	0.5
0.5	239.700	245.56	245.12	Juncu	1.607	1.550	1.199	1.122
0.8	237.570	237.17	246.6	Present	1.627	1.418	1.390	1.110
(a)				(b)				

4 Results

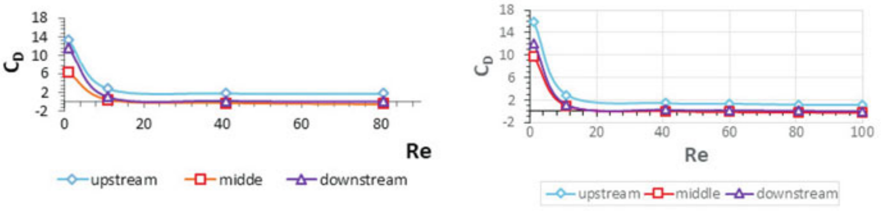
Numerical simulations were done for oblate spheroids with λ = 0.1, 0.2, 0.5, 0.8 and Reynolds numbers from 1 to 100 for different distances between them. For Re = 1 the flow between spheroids with λ = 0.2 lost its Stokes character and clearly visible zones of back flow arise behind the upstream and middle one. Separation of the flow starts from a point close to the rear stagnation one and provokes formation of a ring eddy attached to the rear of the body surfaces. With increasing Re the separation point moves up and the wake in the gap grows and enlarges towards the downstream body. For λ = 0.5 and Re = 1 the back flow area behind the second body is just starting to grow while for spheroids with λ = 0.8 such area is missing. No vortex behind the downstream spheroid is visible for Re = 1 and all values of λ. For higher Re the flow picture changes and at Fig. 2 is shown the flow for $d_1 = d_2 = 1$, λ = 0.2, 0.5 and Re = 60. Let for simplicity the zones of back flow behind upstream, middle and downstream spheroid denote also first, second and third zones respectively. Increasing the fluid velocity, the vortex zones grow and increase their intensity. For distance $d_1 = d_2 = 1$ the third back flow area for spheroids with λ = 0.1, 0.2, and 0.5 appears for Re = 7, 10 and 20 respectively while for spheres at the same distances for Re = 40 [11]. The variation of the drag coefficients with Reynolds number of three spheroids with λ = 0.1 and 0.5, placed at a distance $d_1 = d_2 = 1$ each from another is shown at Fig. 3.

At small Re oblate spheroids with higher λ and placed far from each other show higher total drag coefficients than spheroids with lower λ, i.e. the drag on a sphere is higher than the drag on circular disk that is in accordance with results in [14]. This fact is explained by higher contribution of the drag due to the friction in the total drag, provoked by the increase of the skin area with increase of λ For higher Re this trend is changing as a result of the fast enlargement of the vorticity zones behind the spheroids and the increase in their intensity. The value of Re at which this tendency changes is between 50 and 60. The drag of the middle and downstream spheroid decreases fast and tends to zero quickly with increase of Re and finally accept negative values that means an "attraction effect" exists between the bodies. For small λ, e.g. 0.1 and 0.2 the drag coefficients of the middle and downstream body



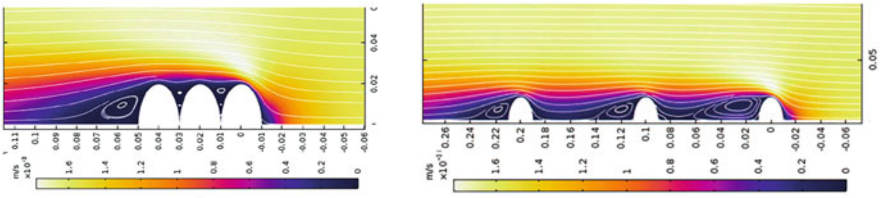
a) $\lambda = 0.2$ b) $\lambda = 0.5$

Fig. 2 Flow past oblate spheroids at distance $d_1 = d_2 = 1$ and Re = 60



a) b)

Fig. 3 Drag coefficient of equal oblate spheroids vs Reynolds number for $d = 1$: (**a**) $\lambda = 0.1$, (**b**) $\lambda = 0.5$



a) b)

Fig. 4 Flow past oblate spheroids for Re = 60, $\lambda = 0.5$: (**a**) $d_1 = d_2 = 0.00001$, (**b**) $d_1 = d_2 = 4$

take negative values for Re about 40 and 80 respectively. The upstream spheroid also changes its drag coefficient up to 11% compared to the single one and depends strongly on the distance between the bodies.

When spheroids with $\lambda = 0.5$ are placed at an extremely small distance from each other, i.e. when they are almost touching and for Re = 50, the total drag of this assemblage is less than the drag of a single spheroid under the same flow parameters. The reason for this effect is the negative values of the drag of the middle and downstream bodies and added to the drag of the first one they reduce the total drag of the assemblage. This effect depends on Re and λ but the most important is the distance between the spheroids. For the mentioned above parameters this effect is valid for distance approx. $d_1 = d_2 = 2$.

At Fig. 4 is shown the flow picture past spheroids being extremely close each to another ($d_1 = d_2 = 0.00001$) and at a distance equal to four large semi-axis ($d_1 = d_2 = 4$) for Re = 60. When the spheroids are very close each to another the eddies in the gaps between them are not clearly expressed. Only a very small recirculating zone exists in the gap between first and second body while behind the middle one vorticity zone has not formed yet. The reason is the narrow gaps where it is difficult for the fluid from the top layers to mix well with the layers beneath. Figure 5 shows the variation of the drag coefficients with the distance between the spheroids: (a) $d_1 = d_2 = d$ and (b) d_1/d_2 i.e. when the middle spheroid changes its location between the upstream and the downstream one. When the distance d is

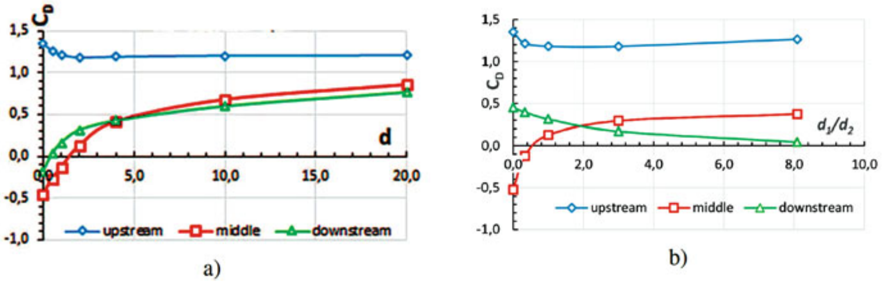


Fig. 5 Variation of the drag coefficients of spheroids with $\lambda = 0.5$ and Re $= 60$: (**a**) C_D vs d, (**b**) C_D vs d_1/d_2

very small, the middle and downstream spheroid feel negative drag and this one of the second body is higher in absolute value than the drag of the downstream body. The negative drag means that the "attractive force" seeks to move the second and third body towards the first one if they were not fixed. Increasing d, the behavior of C_D changes and from a distance of appr. $d = 0.6$ for the downstream spheroid and $d = 1.5$ for the body in the middle, their drag coefficients change from negative to positive. The second spheroid begins to show higher drag than the drag of the third one for $d > 4.5$. Let now change the position of the middle spheroid according to the other two. The goal is to see how this will reflect on the hydrodynamic interaction between particles and their drag coefficients. At Fig. 5b is shown the variation of C_D with ratio d_1/d_2 when the distance between first and third body is fixed to 5, Re $= 60$ and $\lambda = 0.5$. When $d_1/d_2 = 1$ the second spheroid is right in the middle between the other two. Moving it to the first one its drag coefficient rapidly decrease while the drag coefficient of the upstream spheroid increases with appr. 9%. At a distance between upstream and middle spheroid about 1.48 ($d_1/d_2 = 0.55$) the drag coefficient of the middle body takes negative values. If the middle spheroid is moving towards the downstream one ($d_1/d_2 > 1$) the drag on the upstream body keeps almost one and the same value while the drag coefficients of the downstream spheroid decrease but is positive even when the second body is very close to it.

References

1. Stokes, G. G.: Trans. Camb. Phil. Soc. **9**, 8 (1851)
2. Stimson, M., Jefferey, G. B.: Proc. R. Soc. Lond. A **111**, 110–116 (1926)
3. Kim, S.: Phys. Fluids, **30**, 2309–2314 (1987)
4. Tsuji, Y., Morikawa, Y., Terashima, K.: Int. J. Multiphase Flow, **8** (1), 71–82 (1982)
5. Zaprynov, Z. D., Toshev, E. T.: Proc. of the Eighth Int. Conference on Heat Transfer, ASME, New York, 5, 2549–2553 (1986)
6. Kim, I., Elghobashi, S., Sirignano, W. A.: J. Fluid Mech. **246**, 465–488 (1993)
7. Zhu, C., Liang, S. C., Fan, L. S.: Int. J. Multiphase Flow **20** (1), 117–129 (1994)
8. Chen, R.C., Wu, J.L.: Journal of Chem. Eng. Science **55**, 1143–1158, (2000)

9. Schouveiler, L., Brydon, A., Leweke, T., Thompson, M. C.: European Journal of Mechanics B/Fluid **23**, 137–145 (2004)
10. Wilson, H.: Journal of Computational Physics **245**, 302–316 (2013)
11. Kotsev, T.: MATEC Web of Conferences **145**, 03008 (2018)
12. Happel, J., Brenner, H.: Low Reynolds Number Hydrodynamics, Englewood Cliffs, N. J., Prentice Hall Inc. (1965)
13. Pitter, R. L., Pruppacher, H. R., Hamiele, A. E.: J. Atmospheric Sciences **30**, 125–134 (1973).
14. Masliyah, J. H., Epstein, N.: J. Fluid Mech. **44**, part 3, 493–512 (1970)
15. Richter, A., Nikrityuk, P. A.: International Journal of Heat and Mass Transfer **55**, 1343–1354 (2012)
16. Kotsev, T., Zapryanov, Z.: Journal of Theoretical and Applied Mechanics, **18** (3), 36–43 (1987)
17. Ouchene, R.: Phys. Fluids **32**, 073303, 1–11 (2020)
18. Juncu, G.: Int. J. Heat and Mass Transfer **53**, 3483–3494 (2010)

Modelling and Computing the Total Value Adjustment for European Derivatives in a Multi-Currency Setting

Iñigo Arregui, Roberta Simonella, and Carlos Vázquez

Abstract Since the global financial crisis of 2007–2008, different adjustments are considered in the pricing of financial products to incorporate the counterparty risk; the set of these adjustments is referred to as total value adjustment or XVA. In this work we first pose a partial differential equations (PDE) model for pricing the XVA associated to European-like derivatives in multi-currency situations. Moreover, we formulate and solve the XVA pricing problem in terms of expectations to overcome the curse of dimensionality arising in PDEs formulation. Numerical results illustrate the performance of the proposed Monte Carlo algorithms to price best-of-all call options and the sum of put options denominated in different currencies. The second example additionally illustrates the appropriate scaling when the number of stochastic factors (currencies) becomes large.

1 Statement of Partial Differential Equations Model

As a consequence of the financial crisis of 2007–2008, it was clear that the possibility of counterparties default should be taken into account in the pricing of financial derivatives by means of appropriate valuation adjustments, either related to credit (CVA), funding (FVA) or collateral (CollVA), for example. More recently, adjustments related to capital (KVA) or margin (MVA) have been considered. We address the reader to the books [5, 9, 10] and the references therein. In the single currency framework three main approaches have been developed. A first one based on PDEs with seminal references [6, 16], the second one based on expectations started with [4], and the third one based on backward stochastic differential equations [7, 8].

I. Arregui · R. Simonella (✉) · C.Vázquez
CITIC and University of A Coruña, A Coruña, Spain
e-mail: arregui@udc.es; r.simonella@udc.es; carlosv@udc.es

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_41

In the present work we consider a multi-currency setting, following the ideas in [12], where the joint consideration of CVA, FVA, CollVA and repo adjustments are taken into account. We will refer to the set of this adjustments as total value adjustment or XVA. For the additional inclusion of KVA or MVA in the XVA, the ideas in [13, 14] in the single currency case could be considered.

In this section we pose a PDE formulation for the value of a derivative traded in a multi-currency framework, taking into account the total value adjustment to consider possible defaults of the counterparties involved in the deal.

Let $S_t = (S_t^1, \dots, S_t^N)$ be the vector, at time t, of the underlying assets prices $S_t^i, i = 1, \dots, N$, each one of them being denominated in its corresponding *foreign* currency C_i. Moreover, let h_t be the investor's credit spread, and X_t^{D,C_j} (for $j = 0, \dots, N$) the foreign exchange (FX) rate between the *domestic* currency D and C_j, namely the domestic price of one unit of the foreign currency C_j.

The stochastic differential equations (SDEs) governing the evolution of the prices of the underlying assets, the FX rates (see [5]), and the investor's credit spread under the risk neutral probability measure (Q^D) of the domestic market are:

$$dS_t^i = (r^i - q^i)S_t^i \, dt + \sigma^{S^i} S_t^i \, dW_t^{S^i}, \qquad i = 1, \dots, N, \tag{1}$$

$$dX_t^{D,C_j} = (r^D - r^j)X_t^{D,C_j} dt + \sigma^{X^j} X_t^{D,C_j} dW^{X^j}, \quad j = 0, \dots, N, \tag{2}$$

$$dh_t = -\kappa \frac{h_t}{1-R} dt + \sigma^h \, dW_t^h, \tag{3}$$

where r^D and r^i are respectively the risk-free rate in currencies D and C_i, q^i is the dividend paid by S^i, and R is the investor's recovery rate. Moreover, σ^{S^i}, σ^{X^j} and σ^h are the volatility functions of S_t^i, X_t^{D,C_j} and h_t, respectively, while W^{S^i}, W^{X^j} and W^h are correlated Brownian motions. Nevertheless, in the following we consider $\sigma^{X^j} = 0$ in order to have deterministic FX rates.

Next, let J_t^P be the investor's default state at time t, i.e., $J_t^P = 1$ in case of default before or at time t, otherwise $J_t^P = 0$. We use the notation $V_t^D = V^D(t, S_t, h_t, J_t^P)$ for the derivative value at time t from the investor's point of view in domestic currency and $V_t^{RF,D} = V^{RF,D}(t, S_t)$ for the corresponding risk-free derivative price, i.e, traded between two non-defaultable counterparties.

In order to price the derivative, we follow [11, 12] and consider a self-financing portfolio Π that hedges all the risk factors: the market risk due to changes in $S^1, S^2, \dots, S^N$, the investor's spread risk due to changes in h, and the investor's default risk. Moreover, we assume the existence of a collateral account, denominated in currency C_0, composed of a portfolio of bonds R^{C_0} and cash M^{C_0}. We address the reader to [2] for further details.

No arbitrage arguments and the self-financing condition, jointly with the use of Itô's formula for jump-diffusion processes, lead to the following pricing PDE for a European-like derivative with counterparty risk (see [2], for details):

$$\begin{aligned}&\frac{\partial V^D}{\partial t} + \mathcal{L}_{Sh} V^D - f^{H,D} V^D + \frac{h}{1-R}\Delta V^D \\ &\quad = \Big[(r^R + b^{D,C_0} - f^{H,D}) R^{C_0} + (c^D + b^{D,C_0} - f^{H,D}) M^{C_0}\Big] X^{D,C_0}\,, \end{aligned} \tag{4}$$

where $\mathcal{L}_{Sh}$ is the second order differential operator given by

$$\begin{aligned}\mathcal{L}_{Sh} = {} & \frac{1}{2}\sum_{i,k=1}^{N} \rho^{S^i S^k} \sigma^{S^i} \sigma^{S^k} S^i S^k \frac{\partial^2}{\partial S^i \partial S^k} + \frac{1}{2}(\sigma^h)^2 \frac{\partial^2}{\partial h^2} \\ & + \sum_{i=1}^{N} \rho^{S^i h} \sigma^{S^i} \sigma^h S^i \frac{\partial^2}{\partial S^i \partial h} + \sum_{i=1}^{N} (r^i - q^i) S^i \frac{\partial}{\partial S^i} - \frac{\kappa h}{1-R}\frac{\partial}{\partial h}, \end{aligned} \tag{5}$$

and ΔV^D is the variation of V^D upon default defined as $\Delta V^D = RM^+ + M^- - V^D$, with $M(t, S_t, h_t)$ representing the mark-to-market derivative price.

Two possible values for M are usually chosen [6]: either equal to the risky value or to the risk-free value of the derivative. We choose $M = V^D$, so that (4) turns into

$$\frac{\partial V^D}{\partial t} + \mathcal{L}_{Sh} V^D - f V^D = (\bar{r} R^{C_0} + \bar{m} M^{C_0}) X^{D,C_0} + h (V^D)^+, \tag{6}$$

where $\bar{r} = r^R + b^{D,C_0} - f^{H,D}$, $\bar{m} = c^D + b^{D,C_0} - f^{H,D}$ and $f = f^{H,D}$.

Next, we denote by U the XVA price, that can be computed as the difference between the risky derivative value V^D and the risk-free derivative value $V^{RF,D}$. As V^D and $V^{RF,D}$ are both equal to the payoff at maturity T, we have $U(T, S, h) = 0$.

Considering that the risk-free price follows the multidimensional Black-Scholes equation, from (6) we obtain the following nonlinear PDE for the XVA price [2]:

$$\frac{\partial U}{\partial t} + \mathcal{L}_{Sh} U - f U = h\big(V^{RF,D} + U\big)^+ + \big(\bar{r} R^{C_0} + \bar{m} M^{C_0}\big) X^{D,C_0}\,, \tag{7}$$

jointly with the final condition $U(T, S, h) = 0$, where $(t, S, h) \in [0, T) \times (0, +\infty)^N \times (0, +\infty)$. As an alternative, the choice $M = V^{RF,D}$ leads to a linear model [2].

2 Formulation in Terms of Expectations

Since the spatial dimension of (7) increases with the number of currencies, the PDE easily becomes high dimensional in space. Therefore, we propose in this section an alternative expectation-based formulation. In this way, we overcome the so-called *curse of dimensionality*, that affects most of the numerical approaches to solve PDE problems. Thus, we use a Monte Carlo method to approximate expectations in a multidimensional framework, allowing to manage problems that involve more than two stochastic factors.

In order to compute the values of U by using the Monte Carlo method, we apply the nonlinear Feynman-Kac theorem [15], that relates the solution of nonlinear PDEs with the solution of BSDEs. More precisely, Theorem 1.1 in [3] can be applied to formulate (7) in terms of the following nonlinear integral equation:

$$U(t,S,h) = E_t^{Q^D}\Bigg[-\int_t^T e^{-f(u-t)}\Big(h_u \Big(V^{RF,D}(u,S_u) + U(u,S_u,h_u) \Big)^+ \\ + \Big(\bar{r} R_u^{C_0} + \bar{m} M_u^{C_0} \Big) X_u^{D,C_0} \Big) du \mid S_t = S, h_t = h \Bigg]. \tag{8}$$

Analogously to [12], the integrand in the first line of (8) corresponds to CVA+FVA, while the integrand in the second line is related to CollVA and repo adjustment.

In order to compute the XVA given at time $t = 0$, i.e. when the derivative is priced, we numerically solve (8) with a fixed point method and a trapezoidal quadrature formula. Thus, we start from $U^0 = 0$ and recursively compute until convergence:

$$U^{\ell+1}(0,S,h) = E_0^{Q^D}\Bigg[-\int_0^T e^{-fu}\Big(h_u \Big(V^{RF,D}(u,S_u) + U^{\ell}(u,S_u,h_u) \Big)^+ \\ + \Big(\bar{r} R_u^{C_0} + \bar{m} M_u^{C_0} \Big) X_u^{D,C_0} \Big) du \,\Bigg|\, S_0 = S, h_0 = h \Bigg].$$

3 Numerical Results

We now report some results obtained by using the Monte Carlo method for the evaluation of different multi-asset options in the presence of XVA. In all the examples we have considered constant FX rates and maturity T has been set to 6 months. The values of the parameters are specified in Table 1. Moreover, we have used $N_P = 10{,}000$ paths and $N_T = 1000$ time steps. Other test cases are presented in [2].

Table 1 Financial data

$r^1 = 0.30$	$r^2 = 0.24$	$h_0 = 0.20$	$\rho^{S^1S^2} = 0.15$	$R_0^D = 15$	$f = 0.06$
$q^1 = 0.24$	$q^2 = 0.18$	$R_C = 0.30$	$\rho^{S^1h} = 0.40$	$M_0^D = 15$	$\bar{r} = 0.01$
$\sigma^{S^1} = 0.30$	$\sigma^{S^2} = 0.20$	$\kappa = 0.01$	$\rho^{S^2h} = -0.20$		$\bar{m} = 0.02$

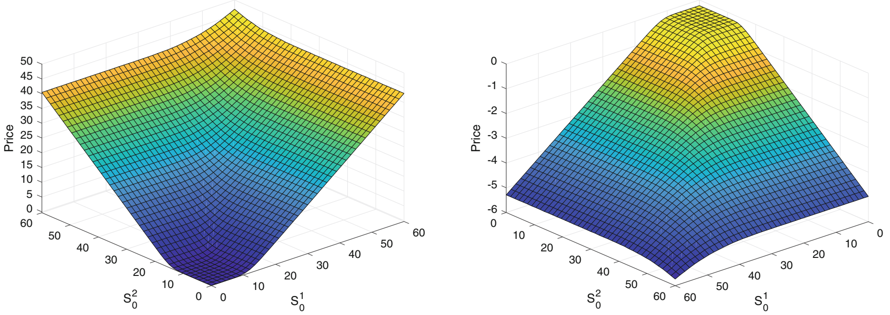


Fig. 1 Best-of-all call option. Price of the risky option (left) and total value adjustment (right)

Table 2 Best-of-all call option. Monte Carlo confidence intervals

		$S^{1,D} = 10$		$S^{1,D} = 14$	
$S^{2,D}$	h	V^D	XVA	V^D	XVA
12	0.10	[0.1200,0.1681]	[−0.2401, −0.2394]	[2.2464,2.3812]	[−0.3591, −0.3542]
12	0.15	[0.1128,0.1609]	[−0.2473, −0.2466]	[2.1785,2.3133]	[−0.4270, −0.4220]
12	0.20	[0.1055,0.1535]	[−0.2547, −0.2539]	[2.1089,2.2437]	[−0.4967, −0.4916]
18	0.10	[3.1003,3.2270]	[−0.4057, −0.3991]	[3.9446,4.0757]	[−0.4521, −0.4439]
18	0.15	[3.0087,3.1354]	[−0.4974, −0.4907]	[3.8293,3.9603]	[−0.5674, −0.5591]
18	0.20	[2.9146,3.0413]	[−0.5915, −0.5847]	[3.7109,3.8420]	[−0.6859, −0.6774]

In the first example we assume the default-free hedger H buys, from a defaultable counterparty C, a European best-of-all call option, the payoff of which is given by:

$$G(t, S^1, S^2) = \max\left((X^{D,C_1} S^1 - K^1)^+, (X^{D,C_2} S^2 - K^2)^+\right), \tag{9}$$

where S^1 and S^2 are two assets respectively written in currencies C_1 and C_2, and K^1, K^2 are the strike values given in the domestic currency D. In our numerical tests, we have chosen $K^1 = 12$ and $K^2 = 15$.

Figure 1 shows the risky option price and XVA, the latter being negative because H asks the counterparty C for a reduction in the price since C may default. Table 2, where the notation $S^{i,D} = X^{D,C_i} S^i$ has been used, shows Monte Carlo 99% confidence intervals for option prices and XVA values for different initial asset prices and investor's credit spread values. Since the credit spread represents the probability of C's default, the XVA value becomes more negative when increasing h.

In the second example we consider that the non-defaultable hedger H buys, from a defaultable counterparty C, a portfolio of N European put options denominated in different currencies, so that the portfolio payoff function is given by:

Table 3 Sum of put options. Monte Carlo confidence intervals and elapsed time

Assets	$V^{RF,D}$	V^{D}	XVA	Time (s)
2	[4.9927, 5.1547]	[4.2446, 4.4104]	[−0.7510, −0.7414]	1.0167
8	[18.7710,19.1090]	[16.5670,16.9110]	[−2.2186, −2.1827]	3.3562
32	[65.9540,66.4760]	[58.7910,59.3200]	[−7.2220, −7.0965]	16.530

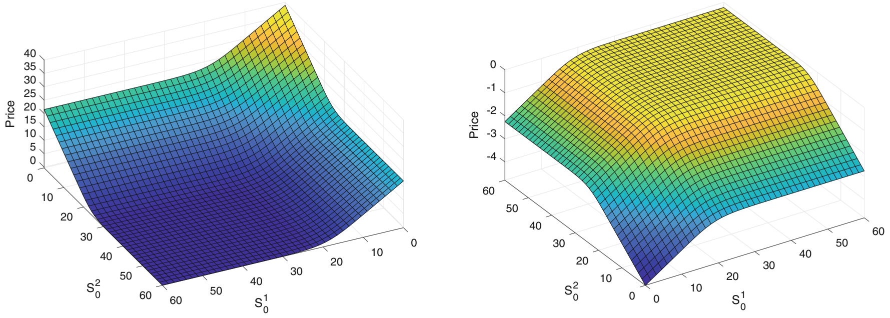


Fig. 2 Sum of put options. Price of the risky option (left) and total value adjustment (right)

$$G(t, S^1, \dots, S^N) = \sum_{i=1}^{N} (K^i - X^{D,C_i} S^i)^+ \,, \tag{10}$$

where S^i $(i = 1, \dots, N)$ are the prices of the underlying assets, respectively written in currencies C_i, while K^i are the respective strike values in the domestic currency D for each put option. Table 3 shows the Monte Carlo 99% confidence intervals for the risk-free, risky and XVA prices for different numbers of underlying assets. Moreover, the elapsed computational time is reported, thus showing a linear increase with the number of assets (stochastic factors). The initial assets prices and the strike values lie in the interval [10, 18].

Finally, we restrict our analysis to the case of the sum of two put options and we set $K^1 = 20$ and $K^2 = 25$. Note that the XVA is negative because the buyer of the derivative H will ask the counterparty C for a reduction in the price due to the potential default of C. As shown in Fig. 2, the XVA becomes more negative when the option is in the money, namely when the asset prices are lower, because H would be more affected by C's default, while the XVA approaches to zero if the asset prices increase, so that the option becomes out of the money. Moreover, the XVA becomes more negative when increasing the number of assets which increases the payoff so that H is more affected by C's default.

4 Conclusions

With the aim of modelling the total value adjustment in a multi-currency setting, we have extended our methodology [1]. Thus, we have stated a nonlinear model and proposed a Monte Carlo method to compute the XVA, that overcomes the curse of dimensionality. We show the suitable performance of the proposed methodology in several examples with European options involving up to 32 underlying assets.

Acknowledgments This work has been partially funded by EU H2020-MSCA-ITN-2018 (ABC-EU-XVA Grant Agreement 813261), Spanish Ministry of Science and Innovation (Grant PID2019-108584RB-I00) and by Galician Government (Grant ED431C2018/033), both including FEDER financial support. Authors also acknowledge the support received from the Centro de Investigación de Galicia (CITIC), funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01.

References

1. Arregui, I., Salvador, B., Vázquez, C., PDE models and numerical methods for total value adjustment in European and American options with counterparty risk. Appl. Math. Comput. 308, 31–53 (2017)
2. Arregui, I., Simonella, R., Vázquez, C., Total value adjustment for European options in a multi-currency setting. Appl. Math Comput. 413, 126647 (2022)
3. Beck, C., Hutzenthaler, M., Jentzen, A., On nonlinear Feynman-Kac formulas for viscosity solutions of semilinear parabolic partial differential equations. (2020). arXiv:2004.03389v2.
4. Brigo, D., Capponi, A., Bilateral counterparty risk valuation with stochastic dynamical models and applications to CDSs, ArXiv preprint, ArXiv:0812.3705 (2009)
5. Brigo, D., Morini, M., Pallavicini, A., Counterparty credit risk, collateral and funding: with pricing cases for all asset classes. The Wiley Finance Series (2013)
6. Burgard, C., Kjaer, M., PDE representations of options with bilateral counterparty risk and funding costs. J. Credit Risk 7, 1–19 (2011)
7. Crépey, S., Bilateral counterparty risk under funding constraints–Part I: pricing, Math. Fin., 25, 1–22 (2015)
8. Crépey, S., Bilateral counterparty risk under funding constraints–Part II: CVA, Math. Fin., 25, 23–50 (2015)
9. Crépey, S., Bielecki. T., Counterparty Risk and Funding: A Tale of Two Puzzles, Chapman and Hall-CRC Press (2014)
10. Gregory, J., Counterparty Credit Risk and Credit Value Adjustment, Wiley Finance (2012)
11. García Muñoz, L.M., CVA, FVA (and DVA?) with stochastic spreads. A feasible replication approach under realistic assumptions. In: MPRA (2013). http://mpra.ub.unimuenchen.de/44568/
12. García Muñoz, L.M., de Lope, F., Palomar, J., Pricing Derivatives in the New Framework: OIS Discounting, CVA, DVA & FVA. In: MPRA (2015). https://mpra.ub.unimuenchen.de/62086
13. Green, A., Kenyon, C., MVA: Initial Margin Valuation Adjustment by Replication and Regression. Risk, 28(5) (2015).
14. Green, A., Kenyon, C., Dennis, C.R., KVA: Capital Valuation Adjustment by Replication. Risk, 27(12) (2014)
15. Pardoux, E., Peng, S., Backward stochastic differential equations and quasilinear parabolic partial differential equations. In: Stochastic partial differential equations and their applications, pp. 20–217. Springer, Berlin (1992)
16. Piterbarg, V., Funding beyond discounting: collateral agreements and derivatives pricing, Risk Magazine, 2, 97–102 (2010)

A Multi-Level Monte-Carlo with FEM for XVA in European Options

Graziana Colonna, Ana M. Ferreiro-Ferreiro, and Carlos Vázquez

Abstract Counterparty credit risk has been recently incorporated in the pricing of financial derivatives by adding different adjustments, the set of which is referred as XVA. In the case of European options to consider stochastic default intensities, instead of constant ones, a three factor model arises. In this work, we have combined a numerical method for solving PDEs with Monte Carlo based techniques, to solve a new hybrid model for XVA pricing. In this way, instead of solving a three dimensional PDEs problem we solve a one dimensional PDE, with two stochastic coefficients coming from the stochastic intensities. More specifically, we propose the use of a Multi-Level Monte Carlo method.

1 Introduction

After Credit Crisis in 2008, unexpected defaults of big companies increased the relevance of counterparty risk in industry and academia. In derivative contracts, counterparty risk refers to the possibility that a counterparty defaults while owing money associated to the contract or while the mark-to-market value of the derivative is positive for the other part of the contract. Many papers and books developed techniques for the valuation of derivatives including counterparty risk by means of valuation adjustments, the set of all of them being referred as total valued adjustment and denoted by XVA. Some particular adjustments included in XVA are:

G. Colonna · A. M. Ferreiro-Ferreiro
CITIC and University of A Coruña, A Coruña, Spain
e-mail: graziana.colonna@udc.es; ana.fferreiro@udc.es

C. Vázquez (✉)
CITIC and University of A Coruña, A Coruña, Spain

ITMATI, Santiago de Compostela, Spain
e-mail: carlosv@udc.es

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_42

- CVA: the cost of hedging counterparty credit risk;
- DVA: the adjustment to a derivative price due to the institution's own default risk;
- FVA: the correction made to the derivative price to account for a funding cost/benefit related to counterparty risk;
- KVA: the cost of holding regulatory capital associated to counterparty risk.

In order to compute the derivative value including the XVA or the price of the XVA, three main approaches are considered in the literature: partial differential equations (PDEs), backward stochastic differential equations (BSDEs) and formulations in terms of expectations. In the PDEs based approach, the spatial dimension of the time dependent PDE is equal to the number of underlying stochastic factors. In many settings, like pricing basket options or interest rate derivatives depending on a large number of forward or swap rates (LIBOR models), the required number of stochastic factors to develop a realistic pricing implies a PDE with high dimension, thus leading to the so called *curse of dimensionality* when numerical methods are addressed. In the present work, in order to overcome the curse of dimensionality, we aim to exploit the combination of Monte Carlo methods and the numerical solution of PDEs with one spatial dimension following the ideas in [5].

2 Modelling with Constant Intensities

Following [3], we consider a derivative contract between two defaultable parties, the hedger (H) and the investor (I). In order to obtain the value of the derivative including counterparty risk, the authors consider a portfolio with four traded assets:

- P^R: default risk-free, zero-coupon bond, with yield r;
- P^H: default risky, zero-recovery, zero-coupon bond of party H, with yield r^H;
- P^I: default risky, zero-recovery, zero-coupon bond of party I, with yield r^I;
- S: underlying asset with no default risk.

Different linear and nonlinear PDEs formulations of the pricing problem can be obtained. The type of PDE depends on the choice of the so called mark to market (MtM) close outs, which is the value of the derivative in case of default. More precisely, let $\hat{V}_t$ be the value of the derivative with counterparty risk (hereafter referred as *risky derivative*), and let V_t be the value of the derivative without counterparty risk (risk free derivative). Possible choices to model the MtM value can be $M_t = \hat{V}_t$ (i.e., equal to the risky derivative value) or $M_t = V_t$ (i.e., equal to the risk free derivative value). In any case, the value of the risky derivative in case of default is:

- $\hat{V}_t = M^+(t, S) + R^I M^-(t, S)$, if the investor I defaults first;
- $\hat{V}_t = M^-(t, S) + R^H M^+(t, S)$, if the hedger H defaults first,

$R^I \in [0, 1]$ and $R^H \in [0, 1]$ being the recovery rates of parties I and H, respectively.

By using dynamic hedging methodology and different versions of Ito lemma, according to the choice of M, two different PDEs arise.

- Non Linear PDE when $M_t = \hat{V}_t$:

$$\begin{cases} \partial_t \hat{V} + \mathcal{A}\hat{V} - r\hat{V} = (1 - R^H)\lambda^H(\hat{V})^- + (1 - R^I)\lambda^I(\hat{V})^+ + s^F(\hat{V})^+, \\ \hat{V}(T, S) = H(S). \end{cases} \tag{1}$$

- Linear PDE when $M_t = V_t$:

$$\begin{cases} \partial_t \hat{V} + \mathcal{A}\hat{V} - (r + \lambda^H + \lambda^I)\hat{V} = -(R^H\lambda^H + \lambda^I)V^- - (R^I\lambda^I + \lambda^H)V^+ + s^F(V)^+, \\ \hat{V}(T, S) = H(S), \end{cases} \tag{2}$$

where we use the differential operator $\mathcal{A} = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2}{\partial S^2} + r_R S \frac{\partial}{\partial S}$ and the notation x^+ and x^- for the positive and negative parts of x. Moreover, r_R is the rate paid in a repurchase agreement, s^F is the funding cost, r^F is the hedger funding rate, λ^H and λ^I denote the constant intensities of default of hedger and investor, respectively. The function H is the pay-off of the derivative in terms of the underlying asset price S.

If U denotes the XVA, then $\hat{V} = V + U$, where V is the price of the risk-free derivative. For a European vanilla option, the Black-Scholes formula provides the value of V. From (1) and (2), we obtain the corresponding PDEs for the XVA price.

- If $M_t = \hat{V}_t$ then U satisfies the nonlinear PDE problem:

$$\begin{cases} \partial_t U + \mathcal{A}U - rU = (1 - R^H)\lambda^H(V + U)^- + (1 - R^I)\lambda^I(V + U)^+ + s^F(V + U)^+, \\ U(T, S) = 0; \end{cases} \tag{3}$$

- If $M_t = V_t$ then U satisfies the linear PDE problem:

$$\begin{cases} \partial_t U + \mathcal{A}U - (r + \lambda^H + \lambda^I)U = (1 - R^H)\lambda^H(V)^- + (1 - R^I)\lambda^I(V)^+ + s^F(V)^+, \\ U(T, S) = 0, \end{cases} \tag{4}$$

Note that in the case of constant intensities of default, there is only one stochastic factor S_t and the spatial dimension of the governing PDE is equal to one. PDEs problems (3) and (4) for constant intensities have been numerically solved in [1], where the method of characteristics (Semi-Lagrangian method) for time discretization is combined with a finite element method (FEM) for the spatial discretization. Additionally, a fixed point iteration is applied to solve the nonlinear PDE.

3 Hybrid Models for Stochastic Intensities

The main objective of the present work is the extension of the previous setting with constant intensities of default to the case with stochastic intensities of default. For this purpose, we pose a hybrid model with three stochastic factors, which is governed by a PDE with two coefficients that are stochastic factors. This approach avoids the alternative consideration of a PDE with three spatial variables, the numerical solution of which is more computationally demanding. Thus, we pose the following linear PDE with one spatial dimension and the two stochastic coefficients:

$$\begin{cases} \partial_t U + \mathcal{A}U - (r + \lambda_t^H + \lambda_t^I)U = (1 - R^H)\lambda_t^H (V)^- + (1 - R^I)\lambda_t^I (V)^+ + s^F (V)^+, \\ U(T, S) = 0, \end{cases}$$

where the stochastic default intensities satisfy the following SDEs:

$$d\lambda_t^I = -\frac{k^I}{1 - R^I}\lambda_t^I dt + \frac{\sigma^I}{1 - R^I} dW_t^I, \quad d\lambda_t^H = -\frac{k^H}{1 - R^H}\lambda_t^H dt + \frac{\sigma^H}{1 - R^H} dW_t^H,$$

with σ^I and σ^H being the volatilities of the intensities of default while k^I and k^H are drift parameters. W_t^I and W_t^H are Brownian motions.

4 Numerical Methods for the Hybrid Model

A first possible naive approach to solve the hybrid model consists in using a crude Monte Carlo (MC) to simulate the paths of the stochastic intensities at the discrete times of the time discretization mesh used for the PDE numerical solution. This method can be sketched as follows:

- Simulate N paths of λ^H and λ^I (i.e., $\lambda^{H,i}$ and $\lambda^{I,i}$, $i = 1, \dots, N$.)
- Solve numerically the (linear or nonlinear) PDE for each path to obtain $\tilde{U}_i$.
- Compute, as solution of the model, the expectation by using Monte Carlo with:

$$\mathbb{E}[\tilde{U}] = \frac{1}{N} \sum_{i=1}^{N} \tilde{U}_i$$

In the present work, we also aim to speed up the Monte Carlo convergence and reduce the variance by using the Multi Level Monte Carlo (MLMC) method presented in [4]. The main ideas of MLMC can be summarized as follows:

If we want to compute the expected value of a process $Q = F(S_t)$, where the process satisfies $dS_t = a(S_t, t)dt + b(S_t, t)dW_t$ and $t \in [0, T]$, we can write:

$$\mathbb{E}[\hat{Q}_L] = \mathbb{E}[\hat{Q}_0] + \sum_{l=1}^{L} \mathbb{E}[\hat{Q}_l - \hat{Q}_{l-1}],$$

where $L > 0$ is a positive integer and $\hat{Q}_l$ is an approximation of Q, estimated on the discretisation of the time interval with the time step $h_l = \frac{T}{M^l}$, M being a positive integer. Let Y_l denote an approximation of $\mathbb{E}[\hat{Q}_l]$, then:

$$Y_L = Y_0 + \sum_{l=1}^{L} Y_l - Y_{l-1}.$$

Therefore, each Y_l is computed with the MC method, using N_l simulations.

5 Numerical Results

We consider an example with a European put option and we compare the case of constant intensities (1-factor model) with a couple of cases with stochastic intensities (2-factor and 3-factor models). We use a linear PDE model, which is numerically solved with the method developed in [1] using a uniform spatial mesh with 1000 nodes and a time step depending on the level in the MLMC method. As MLMC parameters we consider $L = 4$ and $M = 4$, with $N_l = 500$ simulations per level.

First, assuming that $\lambda^I = 0$ we compare the 1-factor and 2-factor models corresponding to the cases λ^H constant and λ^H stochastic, respectively. The values of the parameters are $\sigma = 0.3$, $r = 0.04$, $\lambda_0^H = 0.04$ (constant case and initial intensity in stochastic case), $R^H = 0.4$, $R^I = 0.3$, $k^H \in \{0.1, 0.3, 0.5, 0.7\}$, $\sigma^H = 0.2$, the strike $K = 2$ and the maturity $T = 0.5$. The PDE variables are $t \in [0, T]$ and $S \in [0, 3]$. Next, we compare the 1-factor and 3-factor models, where we additionally consider a stochastic λ^I, with parameters $\lambda_0^I = 0.04$, $k^I \in \{0.1, 0.3, 0.5, 0.7\}$ and $\sigma^I = 0.2$.

In Fig. 1 we show the XVA prices of 1-factor versus 2-factor (left) models and of 1-factor versus 3-factor models (right), illustrating that differences increase for small values of the underlying asset and for larger values of drift coefficients in the stochastic equations governing intensities of default. The XVA is negative as it represents the decrease in the risk free put value due to the probability of default. For small values of the asset, the put option is in the money and one counterparty will be

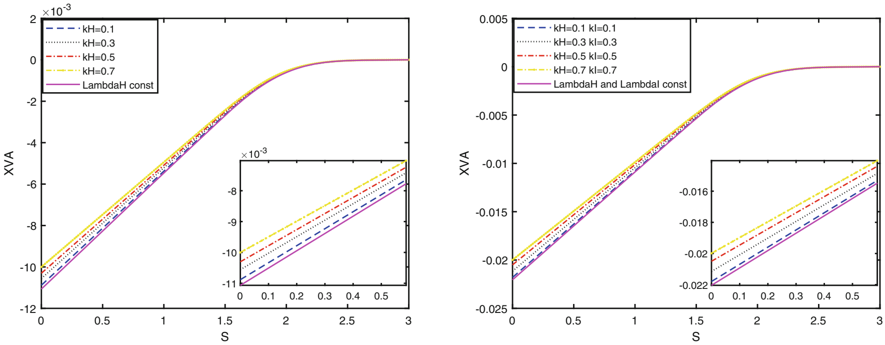


Fig. 1 Comparison 1-factor versus 2-factor models (left) and 1-factor versus 3-factor models (right), for different drifts in stochastic intensities

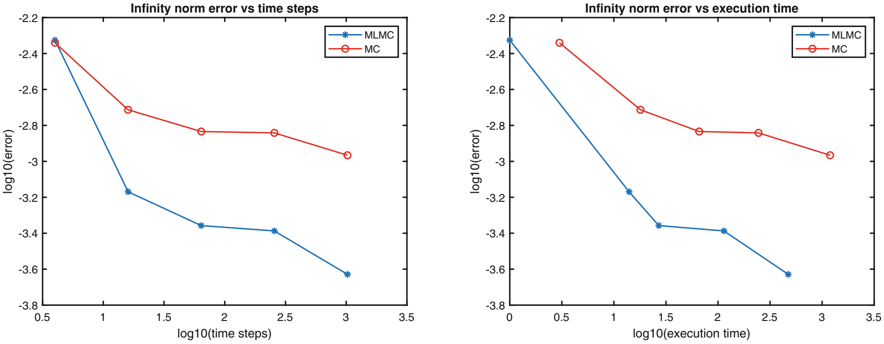


Fig. 2 Comparison of errors in crude Monte Carlo (MC) and Multi Level Monte Carlo (MLMC)

interested in exercising so he/she will be (more) exposed to the other counterparty default. As the exposure has a more negative impact on the put option value for smaller asset values, the XVA becomes more negative and more sensitive to the variation of the drifts of the stochastic intensities of default.

Finally, by using as reference solution the one obtained with MLMC with parameters $L = 5$, $N_l = 2000$, we compare the crude MC and the MLMC for the 3-factors hybrid model. Figure 2 shows the maximum error with respect to time step (left) and computational times (right), clearly illustrating the advantages of MLMC.

6 Conclusions

A hybrid model has been proposed for the case of stochastic intensities of default involving three factors in the evaluation of XVA. The hybrid approach allows to consider PDEs with one spatial dimension and two stochastic coefficients, thus avoiding the solution for PDEs with three spatial dimensions. Multi Level Monte Carlo speeds up the convergences with respect to the use of a crude Monte Carlo numerical methodology. Numerical results illustrate the effect of considering more realistic stochastic intensities of default with respect to constant ones. More details, specially about the numerical examples and their discussion, will appear in [2].

Acknowledgments This work has been partially funded by EU H2020-MSCA-ITN-2018 (ABC-EU-XVA Grant Agreement 813261), Spanish Ministry of Science and Innovation (Grant PID2019-108584RB-I00) and by Galician Government (Grant ED431C2018/033), both including FEDER financial support. Authors also acknowledge the support received from the Centro de Investigación de Galicia (CITIC), funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014–2020 Program), by grant ED431G 2019/01.

References

1. Arregui, I., Salvador, B., Vázquez, C.: PDE models and numerical methods for total value adjustment in European and American options with counterparty risk. Appl. Math. Comput. 308, 31–53 (2017).
2. Colonna, G., Ferreiro-Ferreiro, A.M., García, J.A., Vázquez, C.: A hybrid model based on a Multi Level Monte Carlo FEM for XVA pricing in European options (preprint in preparation).
3. Burgard, C., Kjaer, M.: PDE representations of options with bilateral counterparty risk and funding costs. J. Credit Risk 7, 1–19 (2011).
4. M.B. Giles. Multilevel Monte Carlo methods. Cambridge University Press (2018).
5. T. Lipp, G. Loeper, O. Pironneau. Mixing Monte-Carlo and partial differential equations for pricing options. Chin. Ann. Math. Ser. B 34, 255–76 (2013).

Estimation of Cable Bundle Stiffness Based on Gaussian Process Regression

Lilli Burger, Vanessa Dörlich, Michael Burger, Joachim Linn, and Fabio Schneider

Abstract In modern cars, a huge number of different cables can be found, they are typically combined in hoses and bundles in various different ways. For virtual product development and simulation-based design, it is necessary to know the characteristic physical parameters, like the effective bending or torsion stiffness, of these cable systems. In early stages of the development process as well as for highly customized individual cable configurations, measuring effective stiffness properties is, however, often very challenging. In this contribution, we show results from our current research activities aiming at data-based modeling and estimating effective stiffness parameters for cable bundles. On the basis of an available data set consisting of measured stiffness values for varying cable bundles, the overall goal is to identify a model out of this data, that predicts bundle stiffness values with bundle characteristics as inputs that can be specified without complex measurement efforts. We outline our approach to solve this nonlinear identification task with Gaussian Process (GP) regression. Besides a short introduction to the industrial application area, we demonstrate and illustrate the applicability and prediction quality of Gaussian process regression for this task.

1 Introduction

In modern cars, a huge number of different cables can be found, often combined in cable systems such as bundles or wiring harnesses in various different ways. For virtual product development and simulation-based design, it is necessary to know the characteristic physical parameters, mainly the effective bending or torsion stiffness, of these cable systems. In principle, measuring the effective stiffness values for all cables and assembled cable bundles is possible. However, especially in

L. Burger (✉) · V. Dörlich · M. Burger · J. Linn · F. Schneider
Division Mathematics for Vehicle Engineering, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany
e-mail: lilli.burger@itwm.fraunhofer.de; vanessa.doerlich@itwm.fraunhofer.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_43

early stages of the development process, those measurements are challenging, they might be impractical (due to time and costs) or merely impossible, for instance due to the lack of physical prototypes.

In this contribution, we show results from our current research activities aiming at data based modeling and estimating effective stiffness parameters for cable bundles. On the basis of an available data set consisting of measured stiffness values for varying cable bundles, the overall goal is to identify a model out of this data, that predicts bundle stiffness values with bundle characteristics as inputs that can be specified efficiently without complex measurement efforts. An overview of all available data is given in Sect. 1.2.

We outline our approach to solve this nonlinear identification task with Gaussian Process, which is introduced in Sect. 2 briefly. In Sect. 3, the applicability of Gaussian process regression for this task is illustrated and the performance of this approach in terms of prediction results for different cable bundle types is discussed.

1.1 Simulation of Cable Bundles and Hoses

The demand for software tools that allow a physically correct simulation of slender flexible structures, e.g., cables, cable bundles or hoses, has increased over the last years, especially in automotive industry. A proper framework for modelling of such flexible slender structures is given by the Cosserat rod theory [1], whose geometrically exact kinematics leads to a correct treatment of large rod deformations. The software tool *IPS Cable Simulation* makes use of a discretised geometrically exact rod model [2] enabling a physically correct handling of flexible slender structures in real-time [3, 4]. In this contribution, we focus on the mechanical parameters necessary for modelling the deformations of cable systems, such as single cables and cable bundles. The essential model parameters are the stiffnesses, which give a relation between the deformation measure (e.g. curvature) and the sectional quantity (e.g. moment). A linear elastic constitutive law is assumed, which has proven to be suitable for most practical applications, thus, constant effective stiffness values are sufficient. The deformation modes of main interest are bending and torsion [5]. We restrict our work to an estimation of the bending stiffness $(EI)_b$ of cable bundles, which is assumed to be decoupled from torsional deformation.

1.2 Measurement Campaign

The data required for the estimation process described in the following sections is generated using the MeSOMICS (Measurement System for the Optically Monitored Identification of Cable Stiffnesses) system [6]. It is specifically designed for the measurement of bending and torsional stiffnesses of flexible slender structures, such as cable systems. MeSOMICS uses a bending setup allowing for large

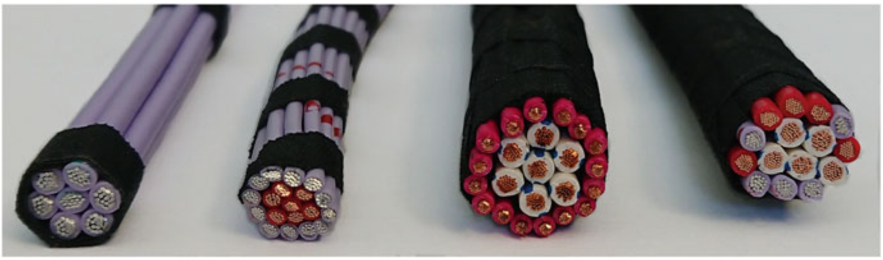


Fig. 1 Examples of different types of bundles with different taping patterns: partly, half and fully, from left to right. The single cables are arranged in layers or randomly (right)

deformations of the specimen and an automated evaluation procedure which enables the generation of a comparatively large database within this work.

The database consists in total of 537 datasets including measurement data of cable bundles and single cables which have been used to assemble the bundles. Aiming at closed packed cross sections to ensure reproducibility, the cables in the bundles are arranged in a regular manner in concentric layers. The database includes measurement data for bundles consisting of one, two or three different types of base cables. In order to reach a high variety, the base cable diameters and types as well as the layer setup are changed systematically, see Fig. 1. For each bundle composition, a textile taping is applied in three different patterns, denoted with partly, half and fully in this work.

2 Gaussian Process Regression

Let $D = \{(\mathbf{X}, y)\}$ be the considered database with input $\mathbf{X} = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N \times d}$ consisting of d different measurement values, $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})^T$, at N measurement points, and corresponding outputs $y = (y_1, \ldots, y_N)^T$.

We assume that there is a mapping f between inputs and outputs. Moreover, we assume independent and normally distributed uncertainty ϵ with standard deviation σ_n^2 in the observed data (e.g. due to measurement errors):

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2). \tag{1}$$

The central task is to choose a model (family) for the mapping f, to train a specific model f based on the observed data D and finally use that model for predicting an output y^* for a suitably chosen (new) input x^* such that $y^* = f(x^*)$. In our approach, we use a stochastic framework, in order to solve this model identification task and, in particular, we choose Gaussian processes (GPs) with mean function m and covariance function k as model structure for the mapping f to be identified,

$$f \sim GP(m, k). \tag{2}$$

A *Gaussian process* is a collection of random variables, for which any finite subset is assumed to have a joint (multivariate) Gaussian distribution. A more detailed discussion of Gaussian processes, especially in the context of machine learning applications, can be found in [7].

In order to specify a GP, $f \sim GP(m, k)$, we have to choose the mean function m and covariance/kernel function k. After that, the key step to understand, how the GP is used for modelling f and for prediction, is to consider the joint distribution of the observed target values y and a requested function value y^*, which is assumed to be Gaussian as well:

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{M_X} \\ m(x_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K} + \sigma_n^2 I & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right), \tag{3}$$

with mean vector $\mathbf{M_x} = [m(x_1), \ldots, m(x_N)]^T$ and a covariance matrix consisting of $\mathbf{K}_{**} = k(x_*, x_*)$, $\mathbf{K}_* = [k(x_*, x_1), \ldots, k(x_*, x_N)]^T$ and $\mathbf{K} = \begin{pmatrix} k(x_1, x_1) & \ldots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \ldots & k(x_N, x_N) \end{pmatrix}$.

Recall that the quantity of interest is y^* as response to x^* (prediction) in the described stochastic framework with a GP, $f \sim GP(m, k)$, as model structure. We consider, consequently, the conditional distribution $p(y^*|x^*, D)$ to derive information about y^*, given x^* and the observed database. According to Eq. (3), this conditional distribution is Gaussian, too, and it is given by

$$\begin{aligned} p(y^*|x^*, D) = \mathcal{N}(m(x_*) + \mathbf{K}_*(\mathbf{K} + \sigma_n^2 I)^{-1}(y - \mathbf{M_X}), \mathbf{K}_{**} \\ - \mathbf{K}_*(\mathbf{K} + \sigma_n^2 I)^{-1}\mathbf{K}_*^T). \end{aligned} \tag{4}$$

Thus, using GPs in this way allows to derive a *predictive distribution* rather than just a point estimate. That is, besides a point estimate, which is, e.g., the mean value of the predictive distribution (4), the prediction automatically provides an uncertainty quantification in terms of the underlying assumed variance.

A Gaussian Process (2) is characterized by its *mean function m* and its *covariance function k*, also referred to as *kernels* [8]. Very often, it is assumed that the GP has zero mean [9], in our application, however, it has shown to be preferable to choose a non-zero mean function, see Sect. 3.1. A generic overview about the mean and covariance functions as well as examples of commonly-used functions is given in [7]. The covariance function describes *similarity* or *nearness* between data points on input level. One of the most common covariance functions is the *squared exponential* covariance function $k(x, x') = \exp(-0.5|x - x'|^2)$. This example illustrates how *similarity* can be interpreted in this case, as $k(x, x')$ is almost equal to one for inputs x and x' very near to each other and decreases with their distance growing.

3 Prediction of Cable Bundle Stiffness Using GP Regression

3.1 Model Setup

We consider one GP per taping option and subdivide the measured data set into a training dataset, consisting of 80% of all available data, and a test dataset, which are the remaining 20% of all data. This results in 134 (partly taped), 131 (half taped), 133 (fully taped) training data points and 34, 33, 33 test data points.

We choose parameters for the single base cables and the bundle itself as well as a nonlinear combination of some of these parameters as predictors and combine them in an input vector $\mathbf{x} = [\sum_c (EI)_c, \sum_c \rho_c, \sum_c r_c, r_b, \sum_c \rho_c \cdot \sum_c r_c / a_c]$, with $(EI)_c$: base cable effective bending stiffness, ρ_c: base cable length density, r_c: base cable radius, a_c: number of cables in the bundle and r_b: bundle radius.

The model output is the effective bundle stiffness, $\mathbf{y} = (EI)_b$. We use a *rational quadratic* covariance function,

$$k(x_i, x_j) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha l^2}\right), r := \left\| x_i - x_j \right\|_2, \tag{5}$$

which is besides the squared exponential covariance function one of the commonly used kernels. The rational quadratic kernel function family includes the squared exponential as limit for $\alpha \rightarrow \infty$ [7]. A systematic analysis of different kernel functions has revealed the rational quadratic to be well suited for our database and use case. Furthermore, our analysis has shown that the quadratic mean function

$$m(x) = \mathbf{h}(x)^T \beta, \mathbf{h}(x) := \left[1, x^{(1)}, \dots, x^{(d)}, \left(x^{(1)}\right)^2, \dots, \left(x^{(d)}\right)^2\right]^T, \tag{6}$$

with parameter vector $\beta = [\beta_1, \dots, \beta_{2d+1}]^T$ is a good choice in this scenario.

After specifying the form of the mean and covariance function, a set of hyper parameters $\theta = [\beta^T, \alpha, l, \sigma_f^2, \dots]$—containing among others the parameter vector β of the mean function and the parameters of the covariance function—have to be derived. This can be done in a Bayesian framework by maximizing an appropriate posterior distribution for the given dataset D. With the additional assumption that the prior distribution of the hyper parameters is approximately constant, i.e., in fact, no prior information, one can obtain a suitable θ by maximizing the log-likelihood $p(y|\mathbf{X}, \theta)$ (e.g., with a standard gradient ascent method). Again, due to the fact that only Gaussian distributions are involved, that log-likelihood can be derived analytically, see [7].

3.2 Prediction Results

With all assumptions and definitions described in Sect. 3.1 we train a model for each taping option and predict the effective cable bundle stiffness with the trained model. In the following, we differentiate between results for all data points (training and test data) and test data only, which gives us the proper prediction quality. We know from simulation studies, that a bundle stiffness deviation of ±50% still leads to sufficient results concerning many applications like, e.g., layout geometry. Thus, we look at the relative prediction error, $e_{rel} = ((EI)_b^{\text{pred}} - (EI)_b^{\text{meas}})((EI)_b^{\text{eas}})^{-1}$, of the effective bending stiffness to assess the prediction quality. To quantify the prediction error more precisely, we consider histograms, which show the ratio of predicted bundle stiffness compared to all data points with a specific relative prediction error. In Fig. 2, such a histogram is depicted for half taped bundles. On the x-axis, the relative error e_{rel} is shown and the bars show the percentage of the predicted bundle stiffness with the according prediction error.

In Table 1, results for all taping options are summarized, subdivided into datasets *all* and *test* and error ranges ±50% and ±20%. It can be observed that, for all taping options, most of the predicted test data (more than 90%) lie within an error range of ±50%, which can be seen as a very good quality. Moreover, considering higher accuracy requirements, we still reach good results. For half taped bundles, 75% of the predicted $(EI)_b$ have a relative error less than ±20% and for fully taped bundles, still more than half of all test predictions (57.6%).

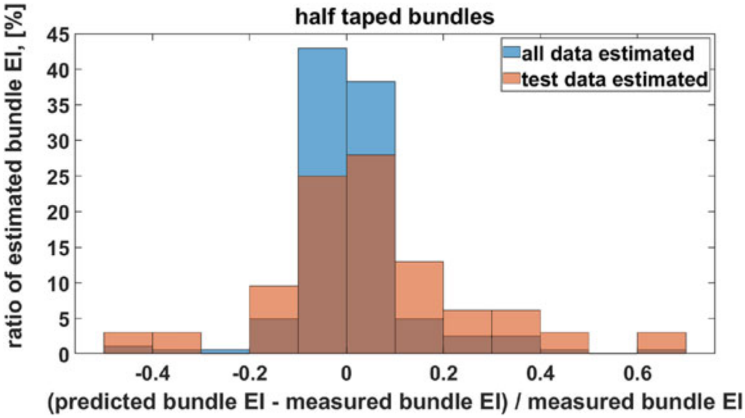


Fig. 2 Histogram of relative prediction error for half taped bundles for all data points (blue) and for test data (orange)

Table 1 Results—ratio of predicted $(EI)_b$ with two different error bounds

Bundle type	All data		Test data	
	$\lvert e_{rel}\rvert \leq 0.5$	$\lvert e_{rel}\rvert \leq 0.2$	$\lvert e_{rel}\rvert \leq 0.5$	$\lvert e_{rel}\rvert \leq 0.2$
Partly	98.2%	90.4%	91.2%	64.7%
Half	99.4%	91.4%	96.9%	75%
Fully	97.5%	85.7%	90.9%	57.6%

References

1. S.S. Antman (2005). Nonlinear Problems of Elasticity. Springer.
2. J. Linn (2020). Discrete Cosserat Rod Kinematics Constructed on the Basis of the Difference Geometry of Framed Curves—Part I: Discrete Cosserat Curves on a Staggered Grid. J. Elast. 139, 177–236.
3. J. Linn, K. Dressler (2017). Discrete Cosserat Rod Models Based on the Difference Geometry of Framed Curves for Interactive Simulation of Flexible Cables. In: Ghezzi L., Hömberg D., Landry C. (eds) Math for the Digital Factory. Mathematics in Industry, vol 27. Springer, Cham.
4. J. Linn, T. Hermansson, F. Andersson, F. Schneider (2017). Kinetic aspects of discrete Cosserat rods based on the difference geometry of framed curves. In ECCOMAS Thematic Conference on Multibody Dynamics, Prague, Czech Republic.
5. V. Dörlich, J. Linn, S. Diebels (2018). Flexible Beam-Like Structures – Experimental Investigation and Modeling of Cables. In Advances in Mechanics of Materials and Structural Analysis (pp. 27–46). Springer, Cham.
6. www.mesomics.eu. Accessed 20 March 2021.
7. C.E. Rasmussen, C.K.I. Williams (2006) Gaussian Processes for Machine Learning. The MIT Press, Cambridge, London. www.GaussianProcess.org/gpml
8. D. Barber (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.
9. M. Ebden (2008). Gaussian Processes for Regression: A Quick Introduction. arXiv:1505.02965

Modeling and Simulation of Pedestrian Interaction with Moving Obstacles Using Particle Method

Parveena Shamim Abdul Salam, Sudarshan Tiwari, and Axel Klar

Abstract Modeling and simulation of pedestrian motion has been an important topic of research in recent years. In this work, we try to understand the dynamics in a shared space of pedestrians and moving obstacles. We consider a social force model coupled with an eikonal equation for pedestrian motion and appropriate kinematic equations for the obstacle motion. Firstly, we attempt to understand how the pedestrians avoid collisions with a passive obstacle. Later we analyze the interaction of pedestrians with a dynamic obstacle having a feedback interaction modeled via a repulsive potential. The hydrodynamic equations are solved using a mesh-free particle method, and the eikonal equation using the fast-marching method. The results reveal the collision avoidance strategies used which are in confirmation with existing studies. The model provides a framework to study pedestrian-vehicular traffic interactions and possibly interactions with automated vehicles in future studies.

1 Introduction

Pedestrian or crowd dynamics has been studied via varied modeling approaches, from microscopic to macroscopic scales. One of the most successful microscopic scale approaches was by modeling pedestrian motion through social or behavioural forces, see [8], which gave insights on self-organisation and collective behaviour of pedestrians like lane formation and bottlenecks. Other agent-based models have also been developed in this scale, for example, in [4]. Macroscopic modeling of the crowd using fluid dynamic equations was introduced by Henderson in [10]. Hughes, followed by others, developed this further via the idea of a potential function in the domain to incorporate more geometric information, see [6, 12]. More macroscopic

P. S. Abdul Salam (✉) · S. Tiwari · A. Klar
Technische Universität Kaiserslautern, Kaiserslautern, Germany
e-mail: parveena@mathematik.uni-kl.de; tiwari@mathematik.uni-kl.de; klar@mathematik.uni-kl.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_44

models are seen in [5, 14]. Elaborate reviews of the different models along with a discussion of their advantages and limitations can be found in [1, 9].

In a social environment, humans encounter stationary or moving obstacles while maneuvering various spaces to reach their destinations. Different voluntary and involuntary strategies are used by humans to avoid collisions in such scenarios. An understanding of when and whether a collision will occur is essential, see [3]. This information forms the basis of collision avoidance models, as seen in [2]. Extensive research on pedestrian interactions with moving obstacles is still limited. In this work, we propose a model to study such interactions in shared spaces. The macroscopic model for pedestrian motion is combined with the proposed kinematic equations of obstacle motion wherein the feedback force terms are modeled via Hughes approach of potential functions obtained through an eikonal equation. For the numerical solutions, an immersed boundary approach is used along with the mesh-free particle method, as seen in [5]. In Sect. 2, we describe the models for pedestrian and obstacle motion. Sect. 3 explains the numerical method briefly. In Sect. 4, we see some results of the numerical simulation for different cases.

2 Models

2.1 Hydrodynamic Model for Pedestrian Motion

The model for pedestrian motion considered is as developed in [5], combining a social force model [8] to a Hughes-type model [12]. The hydrodynamic model equations for the evolution of density ρ and velocity u are:

$$\begin{aligned} &\partial_t \rho + \nabla_x.(\rho u) = 0\,, \\ &\partial_t u + (u.\nabla_x)u = G(x,u,\rho) + \int F(x-y, u(x)-u(y))\,\rho(y)\,dy\,. \end{aligned} \tag{1}$$

These are coupled to the eikonal equation: $f(\rho(x))\,||\nabla\phi|| = 1\,, x \in \Omega$. The force terms in (1), G and F, called the desired acceleration term and the interaction force term, respectively, are defined as:

$$G(\mathbf{x},\mathbf{v},\rho) = \frac{1}{T}\left(-f(\rho)\frac{\nabla\phi(\mathbf{x})}{||\nabla\phi(\mathbf{x})||} - \mathbf{v}\right), \quad F(\mathbf{x},\mathbf{v}) = -\nabla_x U\,, \tag{2}$$

where the potential U with repulsive strength C_r and length scale l_r is given by, $U = C_r \exp\left(-\frac{|x-y|}{l_r}\right)$. We note that a more general F with dependence on both x and v can be used instead of $\nabla_x U$ in (2). The velocity-density relation used is $f(\rho(x)) = u_{\max}(1 - \rho(x)/\rho_{\max})$, where $u_{\max}$ and $\rho_{\max}$ are the maximum velocity and density. We refer to [5] for more details.

2.2 Model for Obstacle Motion

The obstacle motion is governed by kinematic equations for position and velocity. The passive obstacles follow a fixed trajectory defined by the equation: $\frac{dx^O}{dt} = (v_x^O, v_y^O)$ with $v_x^O = -\alpha$ and $v_y^O = A\, cos(\omega t)$ where α is a positive constant for left moving obstacle and A is the amplitude and ω the frequency of oscillatory motion of the obstacle.

The equations for the dynamic obstacle, which interacts with the pedestrians and changes its trajectory or speed, following the convention in pedestrian model, are:

$$\frac{dx_i^O}{dt} = v_i^O\,, \quad \frac{dv_i^O}{dt} = \sum_{j \in N_p} F_O(x_i^O - x_j, v_i^O - v_j) + G_O(x_i^O, v_i^O, \rho_i)\,, \tag{3}$$

which are coupled to the obstacle's eikonal equation, $f_O(\rho(x))\,||\nabla \phi^O|| = 1\,, x \in \Omega$. Here, x_i^O and v_i^O are the position and velocity of the mid-point of the leading edge of the ith obstacle and ρ_i is evaluated at x_i^O by interpolating the density of pedestrians. x_j and v_j are position and velocity of jth neighbour in the list N_p of pedestrians in a circle of radius R centered at x_i^O. Also, we define F_O, G_O and f_O similar to F, G and f as above.

3 Numerical Method

The model equations for pedestrians and obstacle(s) are solved using a mesh-free particle method using least square approximations, see [17]. For this, the hydrodynamic equations in (1) are rewritten in a Lagrangian form as:

$$\begin{aligned} &\frac{dx_i}{dt} = u_i\,, \quad \frac{d\rho_i}{dt} = -\rho_i\, \nabla_x . u_i\,, \\ &\frac{du_i}{dt} = G(x_i, u_i, \delta \star \rho) + \sum_j F(x_i - x_j, u_i - u_j)\, \rho_j\, dV_j\,, \end{aligned} \tag{4}$$

where dV_j is the local area around a neighbouring particle. The kinematic equations of the obstacle(s) in (3) and eikonal equations are coupled to (4) to solve the system completely. An explicit Euler time discretization scheme is used for solving the systems (3) and (4).

The Lagrangian equations are solved on a mesh-free cloud of particles. Furthermore, to solve the eikonal equation, we use an independent structured or unstructured grid on the domain of interest. Information is exchanged between the mesh-free grid and the eikonal grid via interpolation techniques. The eikonal equation is solved by a fast marching method [13, 15]. The boundary conditions of the eikonal equation contain information about the environment, like the position of walls or obstacles. A moving obstacle is treated like an immersed boundary in the eikonal grid, with activation-deactivation of grid points according to the position of the obstacle.

4 Results

Using the numerical method described, we solved the above model equations to analyze the collision-avoidance behaviour of pedestrians and moving passive or dynamic obstacles. We consider a two-dimensional domain of length 100 units and width 50 units for our numerical simulations. The pedestrians are located at the left end of the domain. The right and left boundaries act as exits for the pedestrians and obstacle(s), respectively. Initial pedestrian density is taken as $\rho = 1$ ped/m^2. A fixed time step of 0.002 is used for the explicit time integration scheme.

4.1 Case 1: Passive Obstacle

Passive moving obstacles do not have a feedback interaction with the pedestrians and follow pre-defined trajectory. We considered two different scenarios, pure translation and translation combined with oscillation, and compared with the case of a stationary obstacle. The left and middle subfigures in Fig. 1 show the case where a pedestrian group interacts with a passive obstacle in translation. We observed that when pedestrians interact with a passive obstacle(s), they adjust their path to avoid collision with the obstacle. The path adjustment is made well in advance than the time instance of a head-on collision, using the information available via the eikonal solution. The presence of a moving obstacle slows down the pedestrians, in terms of the time taken to navigate the domain, when compared to their behaviour in the presence of a static obstacle. This implies that the pedestrians exit the domain faster

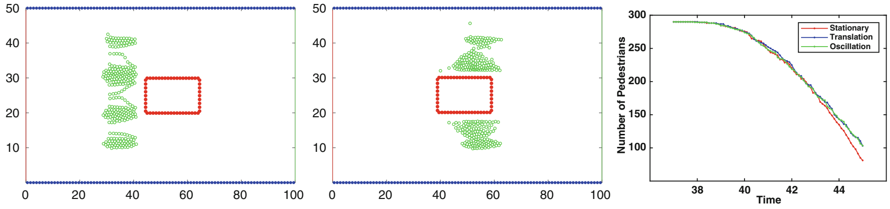


Fig. 1 Pedestrian interaction with a passive moving obstacle shown as red rectangle at time $t = 10s$ (left) and $t = 20s$ (middle). (Right) Number of pedestrians-time graph for the three different cases - stationary obstacle, passive obstacle in translation, passive obstacle in translation and oscillation

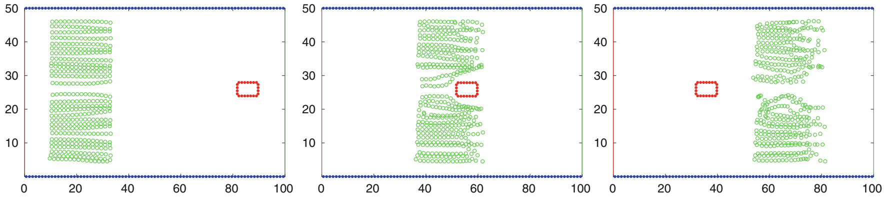


Fig. 2 Pedestrian interaction with a dynamic moving obstacle (red rectangle) at time $t = 5\,\mathrm{s}$ (left), $t = 20\,\mathrm{s}$ (middle) and $t = 30\,\mathrm{s}$ (right). Note that the green markers denote the Lagrangian mesh-free grid points and not the physical pedestrians

in the presence of a stationary obstacle and hence the total density of pedestrians in the domain decreases faster with time as seen in the density-time plot in Fig. 1. This is expected as they have to adjust their path and speed continuously to move forward.

4.2 *Case 2: Dynamic Obstacle*

In the case of a dynamic moving obstacle, both the obstacle and pedestrians actively try to avoid collisions with each other since there is a feedback interaction via the force terms (cf. (1) and (3)). Figure 2 shows a scenario wherein a group of pedestrians interact with a dynamic obstacle. We observe that, though the pedestrians and obstacle(s) undergo path and speed changes, the collision avoidance mechanism is primarily via change of path by pedestrians and change of speed by obstacle(s). Owing to the two-way interactions here, in comparison to one way interaction in the case of passive obstacle, the changes in trajectory of pedestrians is more smoother, continuous and less abrupt. This leads to lesser tendency of having high density of pedestrian crowd near the corners of the leading edge of the obstacle.

5 Conclusion

We have successfully coupled a hydrodynamic model for pedestrian motion with simple kinematic equations for moving obstacles via eikonal equations. Our model satisfactorily replicates the collision-avoidance patterns observed in experimental scenarios like in [11]. But being a macroscopic model, only moderate to high-density scenarios can be studied and it is not possible to analyze microscopic behavioural patterns. We can further study the path and speed changes observed and make quantitative comparisons with other data, for example in [7, 16]. Also, exhaustive studies by changing the size or shape of the obstacle and of the domain can be conducted. We note here that the numerical method used is particularly efficient to employ in complex environments and changes in geometries. For more accurate results, parameters need to be estimated from experimental or real data. Moreover, an extension of the given model to pedestrian-vehicular traffic interactions will be presented in a more elaborate future publication.

Acknowledgments This work is supported by the German research foundation, DFG grant KL 1105/30-1 and grant by the DAAD for Bi-nationally supervised doctoral degree programme.

References

1. Bellomo, N., Dogbe, C.: On the modeling of traffic and crowds: A survey of models, speculations, and perspectives. SIAM Review. **53(3)**, 409–463 (2011).
2. Buisson, J., Galland, S., Gaud, N., Gonçalves, M., Koukam, A.: Real-time collision avoidance for pedestrian and bicyclist simulation: a smooth and predictive approach. Procedia Computer Science. **19**, 815–820 (2013).
3. Cutting, J.E., Vishton, P.M., Braren, P.A.: How we avoid collisions with stationary and moving objects. Psychological Review. **102(4)**, 627 (1995).
4. Degond, P., Appert-Rolland, C., Moussaid, M., Pettré, J., Theraulaz, G.: A hierarchy of heuristic-based models of crowd dynamics. J. Stat. Phys. **152**, 1033–1068 (2013).
5. Etikyala, R., Göttlich, S., Klar, A., Tiwari, S.: Particle methods for pedestrian flow models: From microscopic to nonlocal continuum models. Math. Mod. Meth. Appl. Sci. **24** , 2503–2523 (2014).
6. Di Francesco, M., Markowich, P. A., Pietschmann, J. F., Wolfram, M. T.: On the Hughes model for pedestrian flow: The one-dimensional case. J. Differential Equations. **250**, 1334–1362 (2011).
7. Gao, Y., Chen, T., Luh, P.B., Zhang, H.: Experimental study on pedestrians' collision avoidance. In: Proceeding of the 11th World Congress on Intelligent Control and Automation, pp. 2659–2663. IEEE (2014).
8. Helbing, D. , Molnar, P.: Social force model for pedestrian dynamics, Phys. Rev. E. **51** , 4282–4286 (1995).
9. Helbing, D., Johansson, A.: Pedestrian, Crowd and Evacuation Dynamics, In: Meyers R. (eds) Encyclopedia of Complexity and Systems Science, pp. 6476–6495. Springer, New York (2009).
10. Henderson, L.F.: On the fluid mechanics of human crowd motion. Transportation Research. **8(6)**, 509–515 (1974).
11. Huber, M., Su, Y.H., Krüger, M., Faschian, K., Glasauer, S. and Hermsdörfer, J.: Adjustments of speed and path when avoiding collisions with another pedestrian. PloS one. **9(2)** (2014).
12. Hughes, R. L.: A continuum theory for the flow of pedestrians. Transp. Res. B: Methodol. **36** , 507–535 (2002).
13. Klar, A. , Tiwari, S. , Raghavender, E.: Mesh Free method for Numerical Solution of The Eikonal Equation. In: Proceedings of International workshop on PDE Modelling and Computation, Advances in PDE Modelling and Computation. Ane Books Pvt. Ltd. (2013).
14. Piccoli, B., Tosin, A.: Pedestrian flows in bounded domains with obstacles. Continuum Mech. Thermodyn. **21**, 85–107 (2009).
15. Sethian, J. A.: Fast marching methods. SIAM Rev. **41** 199–235 (1999).
16. Takanashi, H., Kawai, T., Tamura, T., Ota, N.: Prediction of Pedestrian's Walking Route for Moving Obstacles. In: Proceedings of the 5th International Symposium on Future Active Safety Technology toward Zero Accidents (2019).
17. Tiwari, S. , Kuhnert, J.: Finite Pointset Method Based on the Projection Method for Simulations of the Incompressible Navier-Stokes Equation. In: Griebel, M., Schweitzer, M. A. (eds.) Meshfree Methods for Partial Differential Equations. Springer-Verlag (2003).

An Anisotropic Interaction Model for Pedestrian Dynamics with Body Size

Zhomart Turarov

Abstract We propose an extension of the anisotropic interaction model introduced in Totzeck (Kinetic Relat Models, 13(6):1219–1242, 2020) that incorporates the body size of the agents. The rotation of the interaction forces leads to pattern formation. We study the influence of body size on these patterns with the help of numerical simulations.

1 Introduction

The collective behavior of people involving complex mathematical models began relatively recently. Here, the work of the pioneer in this field, Dirk Helbing, is worth mentioning [5, 6, 8]. His work was based on the idea of applying molecular dynamics techniques to crowds [5]. Based on Helbing's model, and looking at different aspects of possible complications of the interaction terms a number of other models have been built, e.g. [7, 12].

Agent-based models can be used in a variety of fields, including transportation stream modeling, epidemiology, sociology, biology, and more [1]. For instance, to describe the swarming of birds, schools of fish, crowd behavior, herd movements, etc. At the same time, it is interesting to study human behavior using agent-based models under various environmental cases. In the work [11], an anisotropic interaction model with collision avoidance is introduced. We extend the model in [11] incorporating body size into pairwise interaction of agents.

This model has been chosen for implementation due to its continuous, multi-agent nature and the ability to vary the accuracy depending on the chosen numerical method. We also can easily select and change the type of interacting forces. To express a behavior aimed at avoiding a collision, we improved the model so that the force vector of pairwise interacting agents rotates.

Z. Turarov (✉)
TU Kaiserslautern, Kaiserslautern, Germany
e-mail: turarov@mathematik.uni-kl.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_45

The model demonstrates several natural behavioral phenomena of pedestrians in motion: move at an individual speed and keep a certain distance from each other [6]. The distance depends on pedestrian density and speed. In evacuation or emergency scenarios, agents push each other and ignore collisions with other pedestrians [9].

In this paper, we represent an experiment involving the bi-directional movement of pedestrians in a corridor. In the following section, we introduce the model. Then, we illustrate the influence of body size on the self-organization of agents in the presented domain.

2 Model

We consider a second order equation of motion with $N \in \mathbb{N}$ agents. Their positions and velocities are denoted by $x_i : [0, T] \to \mathbb{R}^2$ and $v_i : [0, T] \to \mathbb{R}^2$, $i = 1, \ldots, N$. Moreover, the agents are assumed to have a body diameter $d > 0$. This leads to the following interaction dynamics

$$\frac{d}{dt} x_i = v_i, \tag{1a}$$

$$\frac{d}{dt} v_i = \tau \left(w_i - v_i\right) - \frac{1}{N} \sum_{j \neq i} M\left(v_i, v_j\right) K\left(d, x_i, x_j, v_i, v_j\right) \tag{1b}$$

supplemented with the initial conditions $x_i(0) = \mathbf{x}_0,\ \ v_i(0) = \mathbf{v}_0,\ \ i = 1, ..., N$, where $K\left(d, x_i, x_j, v_i, v_j\right) : \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ is a pairwise interaction force between the agents i and j. The rotation matrix $M\left(v_i, v_j\right)$ changes the direction of the interaction force. Applying rotation we get an anisotropic interaction model from the isotropic motion of multi-agent system. The details of the rotation matrix read

$$M\left(v_i, v_j\right) = \begin{pmatrix} \cos\alpha_{ij} & -\sin\alpha_{ij} \\ \sin\alpha_{ij} & \cos\alpha_{ij} \end{pmatrix}, \quad \alpha_{ij} = \begin{cases} \lambda \arccos \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}, & \text{if } v_i \neq 0,\ v_j \neq 0, \\ 0, & \text{else} \end{cases}. \tag{2}$$

In addition, the model includes a relaxation parameter $\tau > 0$ which controls the adaption of the current velocity v_i towards the given desired velocity w_i. The rotation of the force vectors induced by the matrix M models a collision avoidance behaviour of the agents. The direction of the collision avoidance is controlled by the sign of the parameter λ. For $\lambda > 0$ agents move to the right to avoid a collision, for $\lambda < 0$ the movement is directed to the left. See [11] for further details.

The motion of an agent in the model is described by the sum of the forces acting on the agent. By solving the resulting system of differential equations, we can calculate the position and velocity of the agents at any time.

An obstacle can be easily incorporated. For example, the obstacle can be implemented with the help of artificial agents with fixed, predefined positions and artificial velocity vectors pointing outward in the normal vector direction.

2.1 Influence of the Body Size

We extend the collision avoidance model proposed in [11] by introducing an average body size $d > 0$. In more details, we consider the interaction potential

$$U(d, |x_i - x_j|) = R \cdot e^{\frac{d - \|x_i - x_j\|}{r}} - A \cdot e^{\frac{d - \|x_i - x_j\|}{a}},$$

leading to the forces K given by

$$K(d, |x_i - x_j|) = \left(\frac{A}{a} \cdot e^{\frac{d - \|x_i - x_j\|}{a}} - \frac{R}{r} \cdot e^{\frac{d - \|x_i - x_j\|}{r}} \right) \cdot \frac{x_i - x_j}{\|x_i - x_j\|}, \tag{3}$$

where A and R are attraction and repulsion strength, a and r are attraction and repulsion potential ranges.

In the following section, we provide some numerical results that show the influence of the body size on lane formation in the corridor.

3 Numerical Scheme

The initial positions are drawn randomly with uniform distribution in the domain and initial velocities set fixed regarding their direction of motion. Then we solve (1) with a variant of the leap frog scheme [11]. The relaxation terms are solved implicitly and the interaction is solved explicitly as given by

$$\begin{aligned} x_i^{k'} &= x_i^k + \frac{\Delta t}{2} v_i^k, & v_i^{k'} &= (v_i^k + \Delta t \cdot w_i)/(1 + \Delta t), \\ v_i^{k+1} &= v_i^{k'} + \Delta t \cdot \frac{1}{N} \sum_{j \neq i} M(v_i^{k'}, v_j^{k'}) \cdot K(x_i^{k'}, x_j^{k'}), & x_i^{k+1} &= x_i^{k'} + \frac{\Delta t}{2} v_i^{k+1}, \end{aligned} \tag{4}$$

where $i = 1, \ldots, N$, and Δt denotes the step size of the time discretization.

The experiment simulates the movement of two oncoming streams of pedestrians along a spacious corridor. The group of blue agents moves form left to right with desired velocity $w_{\text{blue}} = (0.7, 0)^T$, whereas the red group of agents moves from right to left with desired velocity $w_{\text{red}} = (-0.7, 0)^T$. We consider N_{blue} blue and N_{red} red agents. Hence, the total number of pedestrians in the corridor is

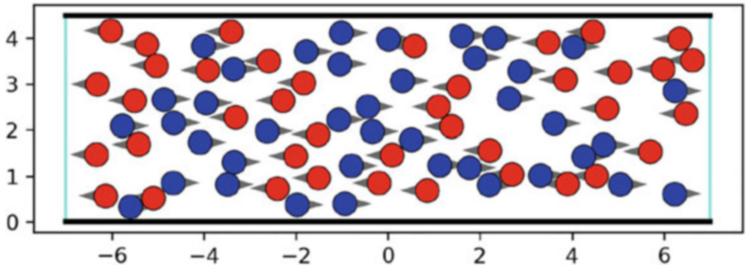


Fig. 1 Initial positions and initial velocity vectors of the agents $N = 80,\ N_{\text{blue}} = 40,\ N_{\text{red}} = 40,$ $d = 0.2$

$N = N_{\text{blue}} + N_{\text{red}}$. The initial positions of the pedestrians $x_i(0) = \mathbf{x}_0,\ i = 1 \dots N$ and their initial velocity vectors $v_i(0) = \mathbf{v}_0,\ i = 1 \dots N$ are presented in Fig. 1.

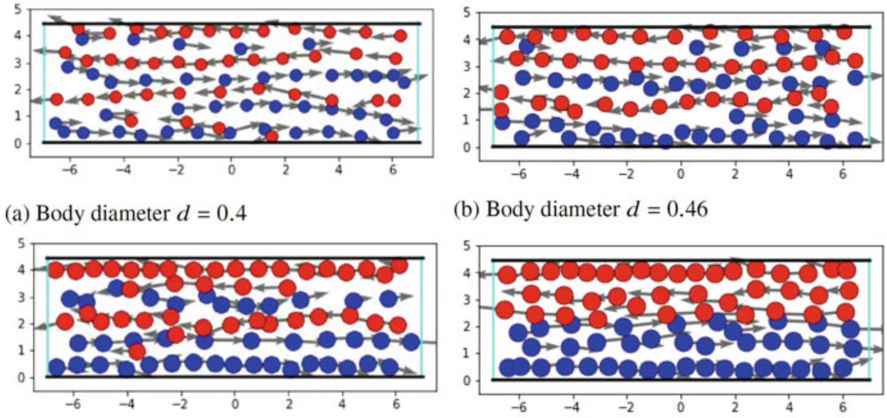
To assure that the pedestrians do not leave the scenario, we add reflective and periodic boundary conditions. In the corridor case the black lines (top and bottom) in Fig. 1 show reflective boundaries. We model the avoidance of wall contact, by reflecting the velocity vector of an agent that would step outside of the domain in the next time step. The light blue lines illustrate periodic boundaries. Blue agents leaving the domain at the boundary on the right, enter again from the left side of the domain. Analogously for the red agents.

3.1 Numerical Study for Different Body Sizes

To analyze the simulation results for different body sizes, we fix values for the force parameters and desired velocities of each pedestrian. The parameters are chosen to satisfy the stability ranges of the interaction force discussed in [4]. In fact, in the range $R/A > 1$ and $r/a < 1$ the interaction force K is repulsive in a short-range, and attractive in a long-range. That allows the distance between pedestrians to be maintained.

Figure 2 shows simulation results of the corridor scenario for different body sizes. The simulation results show, that the formation of lanes in a channel can be reproduced with the help of the rotation anisotropy [11]. The results indicate a relation between the body size and the number of lanes formed. The smaller the body size, the more lanes are obtained. The parameters used for the simulation are reported in the caption of the figure.

In all simulations, we see the formation of so-called traffic lanes. This formation is independent of the choice of the random initial positions and velocities. It is interesting to note that even though every pedestrian is guided by simple rules for movement and interaction, a phenomenon arises that goes beyond the behavior of single pedestrians. Such phenomena of self-organization are manifested in many multi-agent systems [2, 6, 8]. It was reported in many articles concerning the



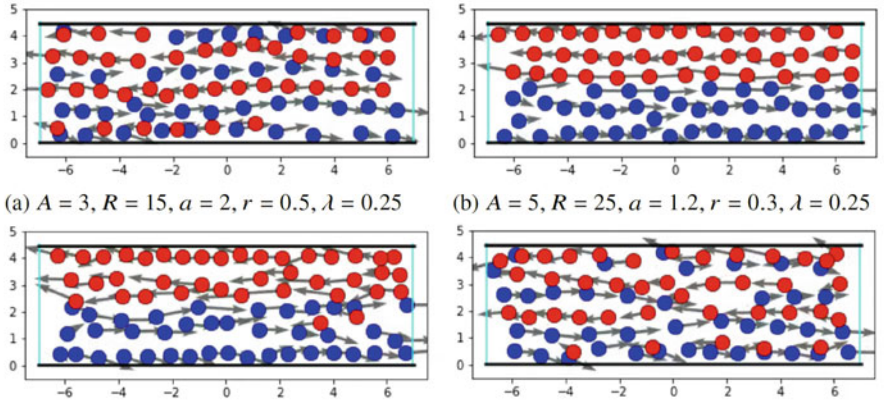
(a) Body diameter $d = 0.4$

(b) Body diameter $d = 0.46$

(c) Body diameter $d = 0.5$

(d) Body diameter $d = 0.6$

Fig. 2 Simulation results in the corridor by different body size of pedestrians at time $T = 35$. On each simulation we fix parameters: $A = 5$, $R = 20$, $a = 2$, $r = 0.5$, $\lambda = 0.25$. Desired velocities for red and blue agents are $w_{\text{red}} = (-0.7, 0)^T$ and $w_{\text{blue}} = (0.7, 0)^T$ respectively. Time step in the Leap-Frog Scheme is $\Delta t = 0.00625$



(a) $A = 3$, $R = 15$, $a = 2$, $r = 0.5$, $\lambda = 0.25$

(b) $A = 5$, $R = 25$, $a = 1.2$, $r = 0.3$, $\lambda = 0.25$

(c) $A = 4$, $R = 26$, $a = 2$, $r = 0.5$, $\lambda = 0.25$

(d) $A = 5$, $R = 25$, $a = 2.4$, $r = 0.6$, $\lambda = 0.25$

Fig. 3 Simulation results in the corridor by different force parameters at time $T = 35$. On each simulation body diameter of the agents are fixed: $d = 0.5$. Time step in the Leap-Frog Scheme is $\Delta t = 0.00625$

movement of pedestrian flows [10, 13], which speaks in favor of the correctness of the proposed model.

It should be mentioned that not only the body size can influence to the number of lanes. In fact, as we can see in Fig. 3 the choice of the corridor width, the number of agents, and attraction and repulsion force parameters can change the formation of lanes as well.

In the future work pattern formation in different scenarios, such as pedestrian flow in crossroads, will be investigated. Besides, based on the data archive of pedestrian dynamics [3], it is possible to calibrate model parameters using an optimal control approach. We create cost functional that minimizes the distance between real and simulated data. With this we get an optimization problem subject to state system (1). It will be interesting to see how well real data can be approximated with the anisotropic model.

References

1. Akopov, A. S., & Beklaryan, L. A. (2012). Simulation of human crowd behavior in extreme situations. International Journal of Pure and Applied Mathematics, 79(1), 121–138.
2. Crociani, L., Vizzari, G., Gorrini, A., & Bandini, S. (2020). Lane Formation Beyond Intuition Towards an Automated Characterization of Lanes in Counter-flows. Collective Dynamics, 5, 25–32.
3. Data archive of experimental data from studies about pedestrian dynamics. https://ped.fz-juelich.de/da/doku.php?id=extdb
4. D'Orsogna, R., Chuang, L., Bertozzi, L., & Chayes., S. Self-Propelled Particles with Soft-Core Interactions: Patterns, Stability, and Collapse (2006) DOI: https://doi.org/10.1103/PhysRevLett.96.104302
5. Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. Physical review E, 51(5), 4282.
6. Helbing, D., & Molnar, P. (1998). Self-organization phenomena in pedestrian crowds. arXiv preprint cond-mat/9806152.
7. Heliövaara, S., Korhonen, T., Hostikka, S., & Ehtamo, H. (2012). Counterflow model for agent-based simulation of crowd dynamics. Building and Environment, 48, 89–100.
8. Helbing, D., Farkas, I. J., Molnar, P., & Vicsek, T. (2002). Simulation of pedestrian crowds in normal and evacuation situations. Pedestrian and evacuation dynamics, 21(2), 21–58.
9. Li, Q., Liu, Y., Kang, Z., Li, K., & Chen, L. (2020). Improved social force model considering conflict avoidance. Chaos: An Interdisciplinary Journal of Nonlinear Science, 30(1), 013129.
10. Sieben, A., Schumann, J., & Seyfried, A. (2017). Collective phenomena in crowds—Where pedestrian dynamics need social psychology. PLoS one, 12(6), e0177328.
11. Totzeck, C. (2020). An anisotropic interaction model for pedestrian dynamics with collision avoidance. Kinetic and related models, 13(6), 1219–1242.
12. Yu, W., & Johansson, A. (2007). Modeling crowd turbulence by many-particle simulations. Physical review E, 76(4), 046105.
13. Zhang, D., Zhu, H., Hostikka, S., & Qiu, S. (2019). Pedestrian dynamics in a heterogeneous bidirectional flow: overtaking behaviour and lane formation. Physica A: Statistical Mechanics and its Applications, 525, 72–84.

Quantitative Characterization of Ductility for Fractographic Analysis

Laury-Hann Brassart, Samy Blusseau, François Willot, Francesco Delloro, Gilles Rolland, Jacques Besson, Anne-Françoise Gourgues-Lorenzon, and Michel Jeandin

Abstract We develop a machine-learning image segmentation pipeline that detects ductile (as opposed to brittle) fracture in fractography images. To demonstrate the validity of our approach, use is made of a set of fractography images representing fracture surfaces from cold-spray deposits. The coatings have been subjected to varying heat treatments in an effort to improve their mechanical properties. These treatments yield markedly different microstructures and result in a wide range of mechanical properties that combine brittle and ductile fracture once the materials undergo rupture. To detect regions of ductile fracture, we propose a simple machine learning network based on a 32-layers U-Net framework and trained on a set of small image patches. These regions most often contain small dimples and differ by the surface roughness. Overall, the machine-learning method shows good predictive capabilities when compared to segmentation by a human expert. Finally, we highlight other possible applications and improvements of the proposed method.

L.-H. Brassart · F. Delloro · J. Besson · A.-F. Gourgues-Lorenzon · M. Jeandin
Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France
e-mail: laury-hann.brassart@mines-paristech.fr; francesco.delloro@mines-paristech.fr; jacques.besson@mines-paristech.fr; anne-francoise.gourgues-lorenzon@mines-paristech.fr; michel-jeandin@mines-paristech.fr

F. Willot (✉)
Mines Paris, PSL Research University, Centre des Matériaux, UMR CNRS 7633, Évry, France

Mines Paris, Centre de Morphologie Mathématique, Fontainebleau, France
e-mail: Francois.Willot@mines-paristech.fr

S. Blusseau
Mines Paris, Centre de Morphologie Mathématique, Fontainebleau, France
e-mail: Samy.Blusseau@mines-paristech.fr

G. Rolland
EDF-Lab Les Renardières, Matériaux et Mécanique des Composants, Moret-sur-Loing Cedex, France
e-mail: gilles.rolland@edf.fr

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_46

1 Introduction

Our ability to quantify the mechanical properties and inner microstructures of materials is based on mechanical testing and models, and commonly requires imaging techniques and analysis [1, 2]. To account for the complex load redistribution of stress within a material, up to crack or pore nucleation, statistical and microstructural aspects are essential [3]. In turn, to perform damage-design, versatile and robust image analysis methods must be developed to quantify and characterize microstructures [4, 5]. The present work is motivated by our need to assess and understand the mechanical response, up to rupture, of certain coatings obtained by cold-spray techniques. Cold-spray deposits generally do not present, in their initial state, satisfying mechanical properties in that respect. Nevertheless, recently-developed techniques that involve heat-treatments allow for microstructural softening [6]. Although these methods have shown promising results, optimizing on the parameters of the thermal treatments requires some level of automatization, as well as sophisticated image analysis to separate ("segment") brittle from ductile fracture regions. A related problem, that of image classification based on failure modes, has been addressed in [7] and [8], using both classical convolutional layers for texture analysis and a modified method based on the adaptive wavelet transform. The present study addresses the problem of image segmentation rather than classification. We make use of a set of fractography images of cold-spray coatings studied in [6] that show both brittle and ductile modes of failure, thus providing an ideal application to the segmentation problem. This short article is divided in three main sections. The material and fractography images are described in Sect. 2. Sections 3 and 4 address the deep-learning architecture and training, and our results. We conclude in Sect. 5.

2 Cold-Spray Deposits and Fracture Surface Images

This work is based on cold-spray projections of 316L-stainless steel from the company "Impact Innovation". Three thermal treatments in a MF7 furnace under air environment, at 600, 800 and 1000 °C are performed, resulting in different coatings. Fracture surfaces for each sample subjected to a three-points bending test (Fig. 1) are observed by scanning electron microscopy (SEM) using a Supra 55 with 15 kV tension, at a 12 mm working distance, a diaphragm of 240 μm and a magnification of ×1500. Secondary electron imaging, sensitive to the surface topography, has been used.

The heat treatment strongly influences the ductile response. Adhesion mechanisms, and particle-particle interfaces, are modified by the heat treatment, leading to a mixed brittle-ductile response [9]. Rupture zones, in particular, are located in particle-particle interfaces in the as-sprayed state and are not seen on the whole fracture surface [6]. As shown in the fractographic analysis in (Fig. 2), fracture

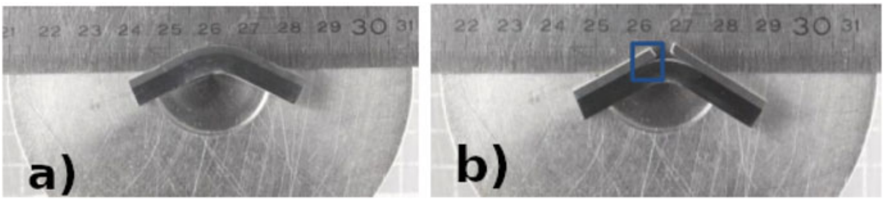


Fig. 1 Three-points bending tests with (**a**) and without (**b**) heat treatment. Heat treatment significantly enhances ductility

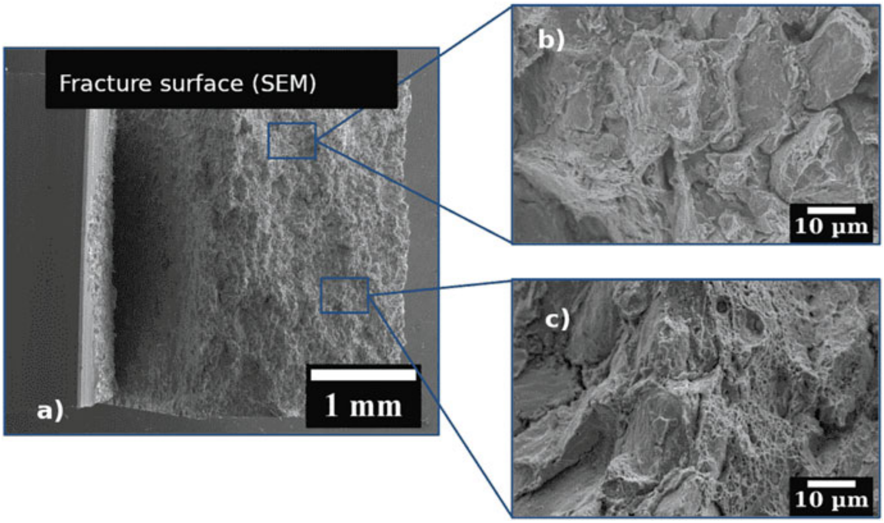


Fig. 2 (**a**) Fractography image showing regions with predominantly brittle (**b**) and ductile (**c**) rupture modes. The bottom zone in (**c**) undergoes maximum bending compared to (**b**)

surfaces display varying contrasts as a consequence of the topography, and ductile regions are constrained along surfaces that rarely exceed 10 μm in size. A signature of these regions is the presence of dimples that are less than 1 μm in diameter. These regions are difficult to detect automatically, yet their texture is characteristic of ductile behavior. In the following, a convolutional neural network is developed to segment ductile regions as a way to provide a more robust method than that of conventional image analysis methods. The proportion of ductile regions in fracture surfaces provides a quantitative indicator of the mechanical response of these coatings, and could be used to correlate microstructure, heat treatment as well as mechanical properties.

3 Machine Learning Method

In order to establish a reference dataset, we first selected 30 images per sample, each containing 1024 × 704 pixels and showing different (disjoint) regions of the fracture surface, with varying contrast and brightness, as specified above. The regions undergoing ductile rupture have been manually annotated by an expert, in all images, making use of a hand-made macro incorporated in the software "ImageJ". These regions are selected by assuming that ductile regions contain dimples and higher roughness. Each of the obtained images are binarized and cropped into 88 patches containing 128 × 64 pixels. Initial images have been subsequently cropped so as to be used during training. Cropping proved necessary to generate a sufficiently large number of images, while reducing the memory required for training. See Fig. 3 for an illustration of the resulting segmentation.

We have selected "U-Net" [10] architecture network for the detection and segmentation of regions with ductile fracture. This architecture, based on a series of filters organized as a "U", allow us to perform four most important tasks: (1) convolution so as to apply one or more filters on the images; (2) nonlinear transformation of images with the rectified linear unit ("ReLu"), so that negative values can be thresholded to zero; (3) pooling so that the image size can be reduced while keeping track of the most important information, including maximum, mean, sum; (4) classification with a fully-connected layer, which links each neuron on the previous layer with a neuron on the following layer, thus allowing one to classify the input pixels according to characteristics highlighted in the previous tasks. In the present study, the number of filters on the first layer is set to 32. During training both training and validation scores are measured. The first one is obtained by measuring the difference between the prediction of the network, made up of regions of ductile failure, and that manually selected by the expert whereas the validation score only takes into account those images that are not part of the training database.

The score is given by the Jaccard distance $J = 1 - |A \cap B|/|A \cup B|$ between two sets A and B, where $| \cdot |$ denotes the set surface. Figure 4 represents the evolution of both scores during training. The scores are plotted with respect to

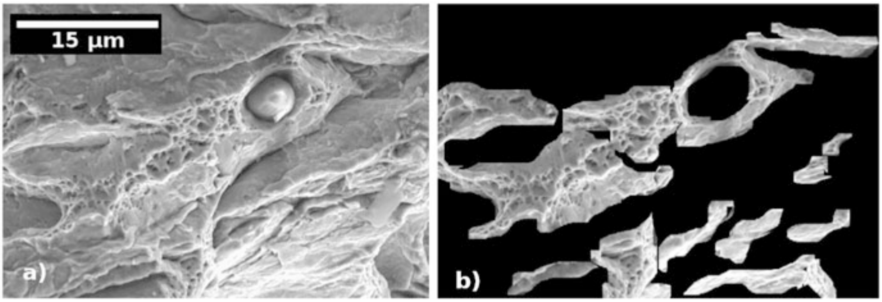


Fig. 3 Annotated fractography images. (**a**) SEM image. (**b**) Selected ductile zones

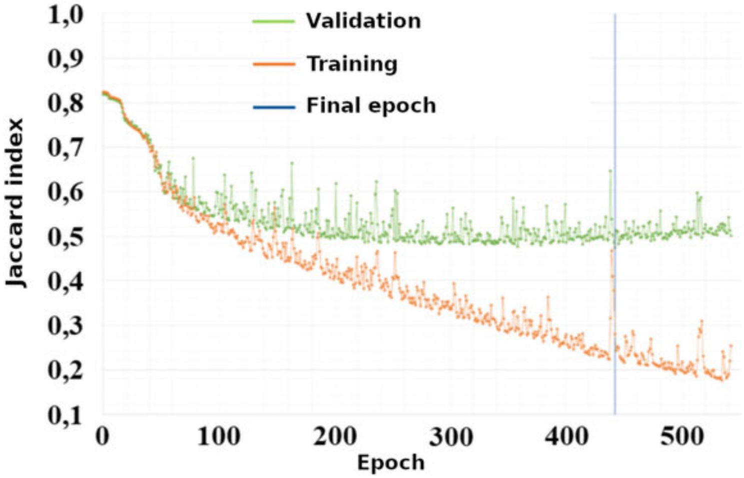


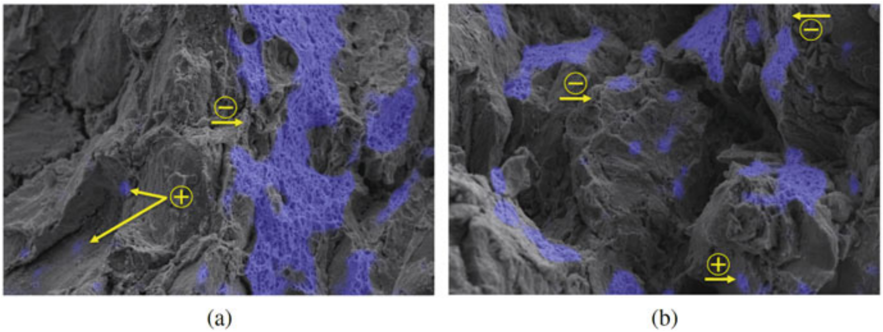
Fig. 4 Validation and training Jaccard indices as a function of the number of epochs during training

epoch numbers. In one epoch, the algorithm uses each patch in the database once. To prevent overfitting, the algorithm is stopped when the score corresponding to the validation dataset ceases to decrease. This is determined by an additional "patience" parameter, set to 100, which prescribes the number of epochs without improvement on the validation dataset that is tolerated during training. The algorithm stopped at epoch 542, whereas the network retained is that corresponding to epoch 442.

4 Results

Two images representative of the network predictions are shown in Fig. 5, as indicated in blue. These images have been obtained by applying the trained U-net network to a set of novel fractography images. We have highlighted in both images zones where the predictions of the network are incorrect: "over-detections" (marked by the symbol $\oplus$), corresponding to brittle regions predicted to be ductile, and "under-detection" (symbol $\ominus$) for ductile zones indicated as brittle.

An error criterion is now defined in order to interpret these results. Output images are thresholded to a value of 200 (out of a maximum of 255), and we compute in each image: (1) the number f of non-detected pixels ("false negative"); (2) the number f' of wrongly-detected pixels ("false positive"); (3) the number t of correct pixels ("true positive"). These statistics allow us to define the precision $p = t/(t + f')$, the recall $r = t/(t + f)$, and score $F = 2pr/(p + r)$, equal to the harmonic mean of p and r. It approaches 1 for predictions close to that of the human expert. The mean of F is about 0.46. This value corresponds to visually satisfactory



(a) (b)

Fig. 5 Two images representing the network predictions for regions undergoing ductile fracture (in blue). Over and under-detections are marked by symbols $\oplus$, $\ominus$, respectively (see text)

segmentation results. Indeed, the F-score is strongly influenced by the exact shape of the ductile zones, which are, in effect, not precisely defined by the expert. Furthermore, we emphasize that the tool is designed to perform comparison between microstructures; accordingly, we aim to rank fractography images by ductile region surfaces, not necessarily to exactly predict the location of each ductile zone.

5 Conclusions and Perspectives

In this work, use has been made of a simple U-Net architecture to segment different fracture mechanisms present in fracture surface images. The images represent complex mixed rupture modes. Ductile rupture is detected by the presence of small dimples seen in the SEM images at various angles from the plane of the image. The network devised in this study shows promising results. However, the machine-learning pipeline tends to detect fewer ductile regions than the human expert, which is conservative but penalizing. Further work is needed to enhance these results, left as outlook: (1) one may improve the image database used for training, and in particular increase the number of images, use larger image patches, or perform data augmentation based on axial symmetries, Gaussian noise and rotations; (2) modify the network architecture, such as the number of filters in the first layer; (3) finally, one may want to adjust or pre-process input images, removing noise and using contrast-enhancing filters. Finally, we emphasize that the simple segmentation method developed in his work can be used in a variety of applications, in particular that of mixed rupture modes, as occurs in the ductile-to-brittle transition of ferritic and bainitic steels [11].

References

1. Tasan C., Hoefnagels J., Ten Horn C., Geers M., Experimental analysis of strain path dependent ductile damage mechanics and forming limits. Mech, Mater. 41(11) (2009), 1264–1276.
2. Dai Q., Sadd M., Parameswaran V., Shukla A., Prediction of damage behaviors in asphalt materials using a micromechanical finite-element model and image analysis. J. Engrg. Mech. 131(7) (2005), 668–677.
3. Bortolussi V., Figliuzzi B., Willot F., Faessel M., Jeandin M., Morphological modeling of cold spray coatings. Image Analysis Stereology 37(2) (2018), 145–158.
4. Abdallah B., Willot F., Jeulin D., Morphological modeling of three-phase microstructures of anode layers using SEM images. J. Microscopy 263(1) (2016), 51–63.
5. Miller K., Akid R., The application of microstructural fracture mechanics to various metal surface states. Materials Science 33(1) (1997), 1–20.
6. Brassart L.-H., Microstructural evolution of 316L stainless steel cold spray coatings under heat treatments; consequences on in-use properties. PhD thesis, Mines Paris (2022).
7. Bastidas-Rodriguez M., Prieto-Ortiz F., Espejo E., Fractographic classification in metallic materials by using computer vision. Engineering Failure Analysis 59 (2016), 237–252.
8. Bastidas-Rodriguez M., Polania L., Gruson A., Prieto-Ortiz F., Deep learning for fractographic classification in metallic materials. Engineering Failure Analysis 113 (2020), 104532.
9. Van Steenkiste T., Smith J., Teets R., Aluminum coatings via kinetic spray with relatively large powder particles. Surface and Coatings Technology 154(2–3) (2002), 237–252.
10. Ronneberger X., Fischer P., Brox T., U-net: Convolutional networks for biomedical image segmentation. Int. Conf. Medical Image Comp. and Comp.-Assist. Interv. (2015), 234–241.
11. Xing R., Chen X., and Yu D., Evolution of impact properties of 16MND5 forgings for nuclear reactor pressure vessel during thermal aging at 500°C. Key Engrg. Mater. 795 (2019), 54–59.

Optimal Control to Facilitate the Development Process of Exoskeletons

Monika Harant, Matthias B. Näf, and Katja Mombaur

Abstract Developing an exoskeleton for the lower back is challenging because the ideal support is not known and may vary across the users. As a result, a series of prototypes and extensive testing is needed to determine a suitable design of such a device. We aim to facilitate the development process of a spinal exoskeleton by optimizing the characteristics of its passive elements, taking into account human-robot interaction. Biomechanical models were created and adapted to the anthropometric and muscular properties of five recorded subjects performing unassisted lifting motions. A dynamic model of the exoskeleton was developed with torque generation consistent with the prototype. Possible configurations of the passive elements are specified by a set of parameters which were determined during optimization by minimizing the simulated human actuation required to perform the recorded motions while wearing the exoskeleton. Comparing optimized and initial setup, a significant improvement in exoskeletal support was achieved for all subjects, while contact forces remained within specified limits to ensure a comfortable usage of the device.

M. Harant (✉)
Mathematics for the Digital Factory, Fraunhofer ITWM, Kaiserslautern, Germany
e-mail: monika.harant@itwm.fraunhofer.de

M. B. Näf
Department of Mechanical Engineering, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: matthias.basil.naf@vub.be

K. Mombaur
Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON, Canada
e-mail: katja.mombaur@uwaterloo.ca

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_47

1 Introduction

In product manufacturing and logistics, industrial workers often face heavy lifting and awkward static postures for long periods of time [11]. These working conditions lead to musculoskeletal disorders, with low-back pain being one of the most common causes for sick leave [1]. Exoskeletons are promising tools to improve the work environment by reducing the muscle activity required to perform certain tasks with high risk of low-back pain, such as lifting, static stooped positions, and overhead manipulation [8]. However, their design process is challenging. Among others, appropriate support profiles and user-acceptable pressure levels are still open research questions and may require studies with a series of prototypes having varying configurations. In [6], the latter issue was tackled by determining pressure thresholds that are still comfortable or free of pain for male and female users. Furthermore, optimal control of biomechanical models has proven to be an effective way to analyze different motions [4, 12], and previous work [3] showed that this setup in combination with an exoskeleton model allows to simulate and optimize its properties and interaction with the user. In this work, we make a step closer to reality by employing a model of an existing prototype [10] instead of a generic one as in [3], and by including interaction force limits based on the findings of [6], which used interfaces similar to the prototype. By optimizing the design for recorded stoop-lifts of five different subjects, we evaluate the current design in terms of its support and applied contact forces during the motion and offer insights on how to improve it further.

Sections 2 and 3 give a brief overview of the experimental data and the applied human and exoskeleton models. The formulation of the optimal control problem to optimize the exoskeleton design for a given lifting motion is described in Sect. 4. The results of the optimization and a short discussion are given in Sects. 5 and 6.

2 Experimental Data

Kinematics, ground reaction forces and forces between box and the subjects' hands of five healthy male subjects (age 21–36 years, weight 60–82 kg, height 1.70–1.82 m) performing stoop-lifts were recorded. A 10 kg heavy box with handles was picked up from a 0.3 m high pedestal placed directly in front of the subjects. Marker positions were recorded at 44 Hz using an Optotrak system (Northern Digital Inc., Canada). Ground reaction forces of the subject and the box were recorded at 1000 Hz with force plates (Kistler Instrumente GmbH, Switzerland) and the forces between hands and handles with uni-directional (vertical) force sensors.

3 Human and Exoskeleton Model

A kinematic analysis shows that the recorded lifting motions that will be used in the optimization are fairly symmetrical. This allows us to reduce the complexity of the system by modeling the human, the exoskeleton, and the box as symmetric rigid (multi)body systems in the sagittal plane. The model of the box matches the one used during the experiment. The human model consists of 11 degrees of freedom (DoF). Due to the symmetric assumption, both arms and legs can be lumped together and the trunk is divided into three parts resulting in the following segments: foot, shank, thigh, pelvis, middle trunk, upper trunk, head, upper arm, and lower arm. For each recorded person, a subject-specific model is created, based on anthropometric measurements taken during the experiment. The human model is actuated by muscle torque generators (MTG) [9], which are adjusted using recorded data as well. Each joint is actuated by two MTG, one for flexion and one for extension. For further information on the human models and on the experimental data please refer to [3].

The exoskeleton model has 9 DoF. The pelvis and the upper trunk module have each 3 DoF (2 prismatic and 1 rotational) and there is a revolute joint at the hip and the thigh interface and a prismatic joint for the slider on the thigh segment. The dynamic parameters are derived from CAD models of the existing prototype. It generates counter torques at the lower back via 3 carbon fiber beams and a passive element (PH) with a nonlinear torque-angle relationship [13] is installed at the hip joint. Mathematical models replicating their behavior are included in the optimization problem and 5 parameters (beam radius, spring pretension and profile dimensions of PH) specify the amount of forces or torques they are generating. For the passive element at the hip joint, the mathematical model of [13] was adopted. For the carbon fiber beam, a polynomial approximation of its deflection is used to obtain the forces generated at the upper trunk module connector.

4 Optimal Control Problem Formulation

The lifting motion of the human model wearing the exoskeleton is separated into 3 phases: The first phase starts when the user stands at rest in an upright position and ends when the user is bent down and makes contact with the box. The second phase covers the force generation to lift the box and ends when it leaves the ground. The last phase ends when the user stands upright again while holding the box. This results in the following 3-phase optimal control problem (OCP):

$$\min_{q,\dot{q},z,\alpha,u,p} \sum_{i=1}^{3} \left(\sum_{n=0}^{N_i} \|W_q(q(t_{i,n}) - q_{i,n}^{REF})\|^2 + \int_{t_i}^{t_{i+1}} \phi(q,\dot{q},z,\alpha,u,p)dt \right) \tag{1}$$

$$s.t. \qquad M(q)\ddot{q} + G_i(q)^T\lambda = \tau(q,\dot{q},z,\alpha,u,p) - C(q,\dot{q}) \tag{2}$$

$$\dot{\alpha} = ((u_m - \alpha_m)/T_m)_{m=1,\ldots,N_m} \tag{3}$$

$$f(q, z, p) = 0 \tag{4}$$

$$g_i(q, \dot{q}, z, \alpha, u, p) \geq 0, \qquad i = 1, \ldots, 3 \tag{5}$$

with q, $\dot{q}$, and $\ddot{q}$ the joint positions, velocities, and accelerations, respectively. The number of shooting nodes of phase i is denoted by N_i. The motion to be tracked is given for time point $t_{i,n}$ by the joint positions $q_{i,n}^{REF}$ and the fitting accuracy is defined by a weighting matrix W_q. The algebraic states z and the system of Eqs. (4) define the state of the beams. The parameters p describe the design of the passive elements of the exoskeleton. The controls u are the neural excitation of the MTG. Eq. (3) are the MTG activation dynamics with activation level α and (de-)activation time constant T. The number of MTG is given by N_m. The equation of motion of the constrained multibody system is given by (2) with M containing the inertia tensors, G_i the constraint Jacobian, and λ unknown force variables. The function C contains the centrifugal, gravitational and Coriolis forces. The generalized forces are denoted by τ consisting of the joint torques and forces generated by the MTG and the exoskeleton. The constraints (5) include, but are not limited to, positional constraints, restrictions on the contact forces between hand and box, box and ground, and foot and ground, regulations on the alignment of human and exoskeleton, and bounds on the states and controls. The objective function (1) consists of a least squares term for tracking the motion and a Lagrange term enforcing the reduction of MTG torques and pelvis contact moment. The last term achieves a balanced force distribution so that the pelvis module does not press into the body. In this case, a tracking term was included as we want to investigate the optimal exoskeleton performance for several recorded lifting motions. For the simulation of the original configuration of the exoskeleton, the parameters p are fixed to the corresponding values. The weighting of the cost function is the same as in case of the design optimization, but the term on the pelvis contact moment was removed. The OCP is discretized using direct multiple shooting and the resulting NLP is solved with SQP and active-set method provided by the toolbox MUSCOD-II [7]. For the rigid multibody dynamics calculations the open-source library RBDL [2] is used.

5 Results

The cost function enforced a high fitting accuracy with avg. joint angle errors within 0.17–0.63° across subjects and stayed the same between design optimization and original configuration so that the reduction in muscle activity comes solely from the support of the exoskeleton and not because of an alteration of the motion. The original design already provides an effective support, but the optimization could increase it significantly for all subjects (Table 1) resulting in a reduction of lumbar moment up to 25.1%, peak lumbar moment up to 18.9%, and hip moment up to 15.1%.

Table 1 Reduction of hip and lumbar moment of the optimized design (initial configuration) with respect to the corresponding human-only-simulation

Subject	Lumbar mom. reduction[a]	Hip mom. reduction[a]	Peak lumbar mom. red.
S1	15.5% (10.4%)	13.6% (14.5%)	18.9% (13.9%)
S2	14.6% (10.2%)	14.3% (13.4%)	14.9% (10.3%)
S3	14.1% (10.7%)	9.2% (9.5%)	15.7% (11.8%)
S4	15.4% (10.1%)	11.3% (5.9%)	14.2% (9.1%)
S5	25.1% (18.0%)	15.1% (13.7%)	17.7% (12.5%)

[a] Reduction in terms of the integrated area under the moment curve

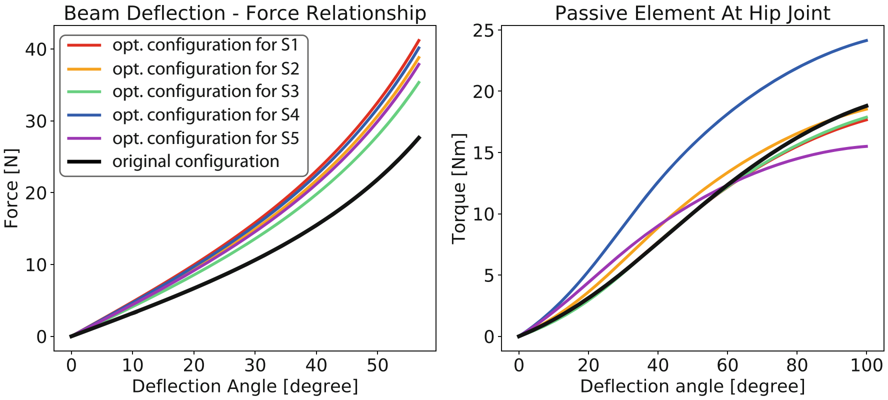


Fig. 1 **Left**: Deflection angle—force relationship of the optimized beam characteristics (colored) and of the original configuration (black). For better comparison, the curves are based on a constant beam length of 40 cm and a force application in a constant direction (90°). **Right:** Deflection angle—torque relationship of optimized (colored) and original (black) configuration of PH

The improved support yields stiffer but quite similar beam characteristics for all subjects (Fig. 1 left). The torque-angle relationships of PH varies more across the subjects (Fig. 1 right). For subject S1, S2, and S3, it stayed close to the original configuration. Significant more torque is produced for S4, who performed the stoop-lift using less hip flexion than the other subjects. For S5, the optimized torque-angle curve became rounder with a lower peak. The forces acting between user and exoskeleton were calculated for 3 contact points: at the back of the pelvis (P), at the front of the chest (C), and at the front of the thigh (T). The limit set on the normal force acting at P (162.4 N) was the most restrictive one for the optimization and was reached across all subjects. The remaining normal forces (C: max. value between 96.0 and 96.7 N; T: max. value between 116.7 and 122.4 N) stayed far away from the prescribed limits (C: 230.3 N and T: 333.4 N). The shear forces (max. value across all subjects: P: 57.8 N; C: 15.9 N; T: 6.9 N) are small throughout the motion. The contact moment acting at P is relatively small during the bending phase (max. value between 6.7 and 17.6 Nm), but is quite high for two subjects at the end of the lift (max. 31.3 and 22.7 Nm) because they arched their back significantly then.

6 Discussion

We simulated an existing prototype for stoop-lifts of five different subjects and optimized the characteristics of its passive elements. This setup not only allows to evaluate different exoskeleton configurations in terms of kinematic structure, torque generation, and contact forces but can also further improve the design by tuning its passive elements. The optimized design yields a significant higher support than the initial setup of [10] across all subjects while contact forces remained within set limits indicating that the torque generation of the prototype can be increased significantly without making it uncomfortable to wear. The optimized support is lower than the values reported in [5], where the prototype was tested with beams of larger diameter than optimized. Reasons for this could be that higher interaction forces occurred during the experiment than were allowed in the optimization, that the subjects altered their behavior slightly, or that there is a discrepancy in the simulated interaction forces because of unaccounted movement of the exoskeleton with respect to the user. In summary, this work highlights that the presented approach can identify untapped potential in the support, but also stresses the need for an accurate simulation of the human-robot interaction as well as sound contact force limits.

Acknowledgments This work was funded by the European Commissions within the H2020 project SPEXOR (GA 687662).

References

1. Factsheet 10 - Work-related Low Back Disorders. European Agency for Safety and Health at Work, Belgium 4 (2000), 41.
2. M.L. Felis, RBDL: An efficient rigid-body dynamics library using recursive algorithms. Autonomous Robots (2016), 1–17.
3. M. Harant, M. Millard, N. Šarabon, K. Mombaur, Cost function evaluation for optimizing design and actuation of an active exoskeleton to ergonomically assist lifting motions. In IEEE/RAS Int. Conf. Humanoid Robots (Humanoids) (2019), 186–193.
4. K.A. Inkol, C. Brown, W. McNally et al., Muscle torque generators in multibody dynamic simulations of optimal sports performance. Multibody System Dynamics 50 (2020), 435–452.
5. A. S. Koopman, M. Näf, S. J. Baltrusch et al., Biomechanical evaluation of a new passive back support exoskeleton, Journal of Biomechanics (2020), 105.
6. Ž. Kozinc, J. Babič, N. Šarabon., Human pressure tolerance and effects of different padding materials with implications for development of exoskeletons and similar devices. Applied Ergonomics 93 (2021), 103379.
7. D.B. Leineweber, A. Schäfer, H.G. Bock, J.P. Schlöder., An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization: Part II: Software aspects and applications, Computers & Chemical Engineering, Elsevier. 27(2) (2003), 167–174.
8. M.P. de Looze, T. Bosch, F. Krause et al., Exoskeletons for industrial application and their potential effects on physical work load. Ergonomics, Taylor & Francis 59(5) (2016), 671–681.

9. M. Millard, A.L. Emonds, M. Harant, K. Mombaur, A reduced muscle model and planar musculoskeletal model fit for synthesis of whole body movements. Journal of Biomechanics 89 (2019), 11–20.
10. M.B. Näf, A.S. Koopman, S. Baltrusch et al., Passive Back Support Exoskeleton Improves Range of Motion Using Flexible Beams. Frontiers in Robotics and AI 5 (2018), 72.
11. A. Parent-Thirion et al., Fifth European Working Conditions Survey. Luxembourg: Publications Office of the European Union, 2012.
12. K. Stein, K. Mombaur., Whole-Body Dynamic Analysis of Challenging Slackline Jumping. Applied Sciences 10(3) (2020), 1094.
13. B. Vanderborght, N.G. Tsagarakis, C. Semini et al., MACCEPA 2.0: Adjustable compliant actuator with stiffening characteristic for energy efficient hopping, IEEE Int. Conf. on Robotics and Automation. 2009:544–549.

Vanadium Redox Flow Batteries: Asymptotics and Numerics

Michael Vynnycky and Milton Assunção

Abstract Modern demands for increasingly efficient renewable energy delivery have generated substantial interest in vanadium redox flow batteries (VRFBs) as an energy storage technology, with mathematical modelling and numerical simulation playing an increasingly important role in their development. Although the overwhelming majority of work in this area tends to involve time-demanding computation, this contribution summarizes our own recent activities in deriving asymptotically reduced versions of the multi-dimensional transient models that are normally used to describe the operation of a VRFB. We find that our models are able to predict the charge-discharge curve and the state of charge of the VRFB as accurately as two-dimensional transient models, but typically at around 1/250th of the computational cost.

1 Introduction

Current demand for increasingly efficient renewable energy delivery has generated substantial interest in vanadium redox flow batteries (VRFBs) as an energy storage technology. VRFBs have numerous potential applications: load levelling and peak shaving, uninterruptible power supplies, emergency backup and facilitation of wind and photovoltaic energy delivery [1, 2].

A VRFB consists of an assembly of cells, typically referred to as a stack; one such cell is shown in Fig. 1. It is composed of positive and negative flow-through electrodes, typically made of porous carbon felt, that are separated by a proton exchange membrane that consists of charged molecules: the mobile protons that pass through it and fixed sites of negative charge. During operation, vanadium-based electrolytes are pumped through the electrodes; the electrolyte in the positive

M. Vynnycky (✉) · M. Assunção
Department of Mathematics and Statistics, University of Limerick, Mathematics Applications Consortium for Science and Industry (MACSI), Limerick, Ireland
e-mail: michael.vynnycky@ul.ie; milton.o.assuncao@ul.ie

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_48

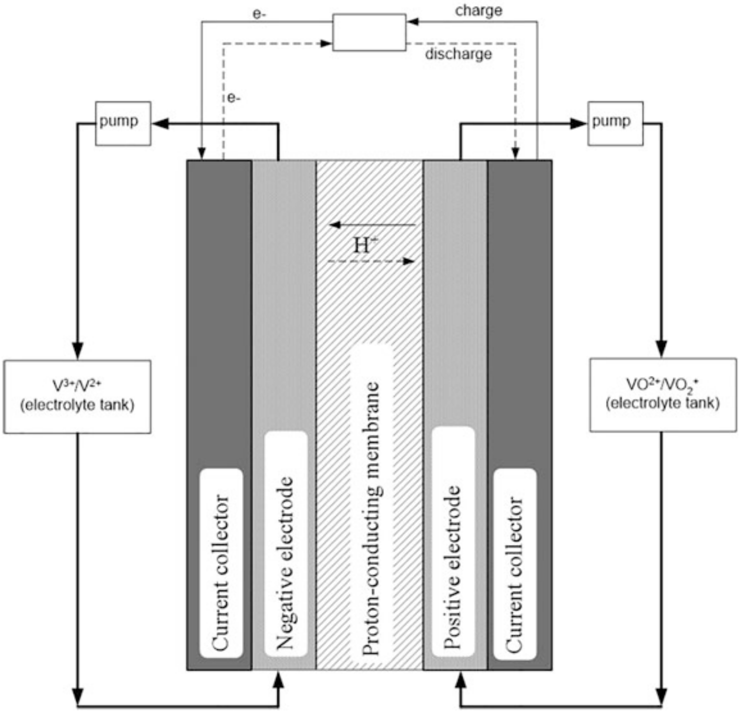


Fig. 1 A schematic of the overall operation of a vanadium redox flow battery

electrode, vanadyl sulphate ($VOSO_4$), contains VO_2^+ and VO^{2+} ions, whilst that in the negative electrode, vanadium sulphate ($V_2(SO_4)_3$), contains V^{2+} and V^{3+} ions. In addition, both electrodes are connected to pumps and storage tanks, meaning that very large electrolyte volumes can be circulated through the cell. During charging, the VO^{2+} ions in the positive electrode are reduced to VO_2^+ ions, and electrons exit from the positive terminal of the cell via a current collector that bounds the electrode on the side opposite to that of the membrane. Similarly, in the negative electrode, electrons enter via another current collector, reducing the V^{3+} ions to V^{2+} ions; during discharge, the reverse process, also known as oxidation, occurs. Charging and discharging can be written as

$$V^{3+} + e^- \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftharpoons}} V^{2+}\text{at the negative electrode,} \tag{1}$$

$$VO^{2+} + H_2O \underset{\text{discharge}}{\overset{\text{charge}}{\rightleftharpoons}} VO_2^+ + e^- + 2H^+\text{at the positive electrode.} \tag{2}$$

Typically, each cell in a VRFB operates at a nominal voltage in the interval 1.15–1.55 V and at a temperature of around 30 °C.

Mathematical modelling and numerical simulation have recently come to play an increasingly important role in VRFB research and development; for recent reviews in all aspects of VRFBs, and modelling in particular, see [1, 2], respectively. In general, the models in question consist of a system of two- or three-dimensional time-dependent partial differential equations (PDEs) that describe the transient mass, momentum and charge transport that occur in the processes mentioned above, and invariably require numerical solution. However, Vynnycky [3] suggested the use of asymptotic methods to reduce these full models; subsequently [4, 5], it was shown via numerical simulations that the governing equations were in fact quasi-steady in nature and that even a zero-dimensional model was able to reproduce the charge-discharge curves.

In this contribution, we give an overview of the subsequent work of Vynnycky and Assunção [6, 7], who formally demonstrated that a standard and often-used VRFB model could be reduced asymptotically to give a much simpler set of equations which had a quasi-analytical solution. This was done for two different scenarios. One includes the dissociation of sulphuric acid (H_2SO_4), a two-step reaction in which the first dissociation step,

$$H_2SO_4 \rightarrow H^+ + HSO_4^-, \tag{3}$$

is assumed to be complete, meaning that H_2SO_4 has completely dissociated into its ions, whereas the second, given by

$$HSO_4^- \rightarrow H^+ + SO_4^{2-}, \tag{4}$$

is incomplete, meaning that all ions shown in Eq. (4) are present. The second scenario excludes the dissociation of H_2SO_4. The reason for focusing on this feature, rather than the numerous others which are believed to be subsidiary to reactions (1) and (2), is that the first scenario is included by default in the VRFB starting model available in the commercially available finite element software Comsol Multiphysics; thus, our motivation was to determine how necessary this feature actually was. A complete list of model assumptions is given in [7, p. 175].

2 Mathematical Modelling

2.1 *Full Model*

The full model in dimensional form and its asymptotic reduction is rather lengthy, and the complete details are given in [6, 7]; here, we give only a qualitative description.

For each of the electrodes, we obtain five time-dependent convection-diffusion-migration-reaction equations, which account for the concentrations of the five reacting ionic species in the electrolyte: H^+, HSO_4^-, SO_4^{2-}, V^{2+}, V^{3+} at the negative electrode; H^+, HSO_4^-, SO_4^{2-}, VO_2^+, VO^{2+} at the positive electrode. However, because of migration, there is an additional dependent variable: the ionic potential. Consequently, a further equation is required in order to fully specify the system; this comes via an electroneutrality condition, which is written as

$$\sum_i z_i c_i = 0, \tag{5}$$

where z_i is the charge number for ionic species i, c_i is concentration for this species, and the summation is taken over all of the ions present in each of the electrodes. Moreover, earlier work has shown that convection is adequately described by simply assuming a plug flow [8]. However, the description so far relates only to the electrolyte phase; through the solid matrix, there is electron transfer and this is described, on using Ohm's law, via a Poisson-type equation for the electronic potential. In particular, the source term for this Poisson-type equation describes reactions (1) and (2) via Butler-Volmer expressions which contain the difference of the ionic and electronic potentials. As for the membrane which separates the two electrodes, only H^+ is assumed to be present, and this fact, on again using Ohm's law, simply leads to Laplace's equation, albeit for the ionic potential.

The key boundary conditions are at the electrode inlets, which are located at the bottom of Fig. 1. An unusual feature is that we cannot just prescribe the concentrations there, since the electrolyte is being recycled. This leads to five first-order ordinary differential equations (ODEs) at each inlet, one for each species concentration, with time as the independent variable; furthermore, these ODEs each contain a term related to the concentration at the outlet, located at the top of Fig. 1, since the electrolyte, on exiting the outlet, is fed back to the storage tank and then to the inlet.

In summary, the full model consists of 11 coupled time-dependent nonlinear PDEs and two algebraic relations, in addition to 10 ODEs at the inlets.

2.2 Asymptotically Reduced Model

As explained in [6, 7], the asymptotic reduction makes use of the fact that the geometry is slender. However, on assessing the numerical values of the model parameters, it is then also noted that there are concentration boundary layers on the porous electrode side of the electrode/membrane interfaces. If acid dissociation is neglected, the boundary layer is of non-dimensional width $Pe^{-1/2}$, where Pe is the ratio of the effects of convection and diffusion, with $Pe \approx 230$, i.e. $Pe \gg 1$. On

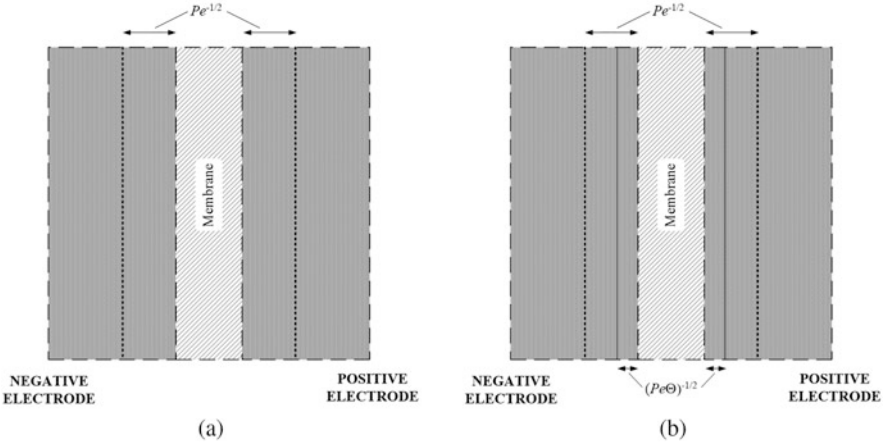
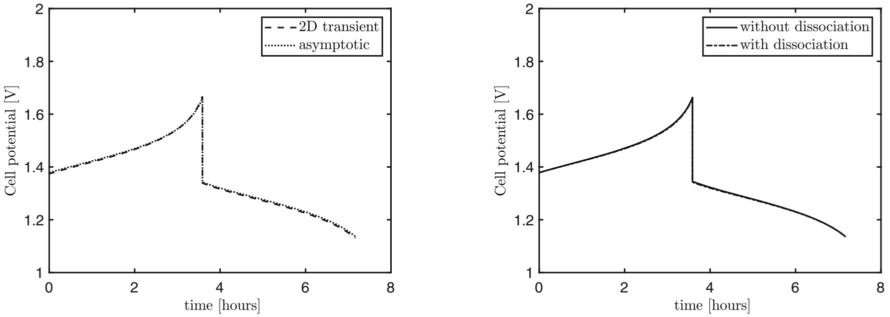


Fig. 2 A schematic for the overall asymptotic structure for the model: (**a**) without H_2SO_4 dissociation; (**b**) with H_2SO_4 dissociation

the other hand, if acid dissociation is accounted for, there is a nested boundary-layer structure, with a layer of width $Pe^{-1/2}$ housing an inner layer of width $(Pe\Theta)^{-1/2}$, where Θ is the ratio of the effects of electrochemical reaction and convection, with $\Theta \approx 166$, i.e. $\Theta \gg 1$. These structures are depicted in Fig. 2, and the result is considerably different to that postulated in [9]; this suggests that the structure was not properly understood earlier.

Further analysis then indicates that, under the operating conditions of constant current density, the inlet concentrations can be determined analytically. In fact, all ten of them are just linear profiles of time, with the profiles for V^{2+}, V^{3+}, VO_2^+ and VO^{2+} when acid dissociation is neglected being identical to the corresponding ones when acid dissociation is included; the profiles for H^+, HSO_4^-, SO_4^{2-} do differ for the two cases [7, Fig. 7], however, as a direct consequence of the acid dissociation. Interestingly, earlier models [10, 11] had assumed that the inlet concentrations were linear, but [6, 7] were the first to show why this is the case.

Subsequently, the remaining task involves the numerical solution of four coupled second-order nonlinear ODEs, with the independent variable being the spatial coordinate across the VRFB, i.e. in the horizontal direction in Fig. 1, and the dependent variables being the ionic and electronic potentials in the negative and positive electrodes. Also, we point out that, by this stage of reduction, the governing equations are quasi-steady, since there are no time derivatives, although time-dependency enters via coefficients that contain the time-linear concentration profiles.



(a)

(b)

Fig. 3 Charge-discharge curve at a typical current density (400 A m^{-2}): (**a**) as predicted by the 2D transient and the asymptotic models for the case with acid dissociation; (**b**) as predicted by the asymptotic model, including and excluding acid dissociation. Reproduced from [7]

3 Results and Conclusions

Figure 3a shows the cell potential as a function of time, often referred to as the charge-discharge curve, at a current density of 400 A m^{-2}, as predicted by the 2D transient and the asymptotic models for the case with acid dissociation; as is evident, the agreement is very good, indicating that the asymptotically reduced model successfully captures the features of the full model. The same result was also obtained earlier in the model without acid dissociation [6, Fig. 4]. Similarly, Fig. 3b compares the charge-discharge curves obtained via the asymptotic approach, including and excluding acid dissociation; as can be seen, the two curves are literally on top of each other, indicating that the acid dissociation has no effect.

Finally, we point out that the numerical solution of the fully reduced asymptotic model was found to require around 250 times less computational time than that of the original 2D transient model, both with and without acid dissociation; the exact details as regards the latter can be found in [6, Table 6]. More significantly, this suggests that other effects that are believed to be present during VRFB operation, such as oxygen and hydrogen evolution, heat transfer and vanadium ion transport across the membrane, can be incorporated into the model, without necessarily increasing the computational time required to solve the model equations.

References

1. Lourenssen, K., Williams, J., Ahmadpour, F.. Clemmer, R., Tasnim, S.: Vanadium redox flow batteries: a comprehensive review. J. Energy Storage **25**, Article no. 100844 (2019)
2. Aramendia, I., Fernandez-Gamiz, U., Martinez-San-Vicente, A., Zulueta, E., Lopez-Guede, J. M.: Vanadium redox flow batteries: a review oriented to fluid-dynamic optimization. Energies **14**, Article no. 176 (2021)

3. Vynnycky, M.: Analysis of a model for the operation of a vanadium redox battery. Energy **36**, 2242–2256 (2011)
4. Sharma, A.K., Vynnycky, M., Ling, C.Y., Birgersson, E. Han, M.: The quasi-steady state of all-vanadium redox flow batteries: a scale analysis. Electrochimica Acta **147**, 657–662 (2014)
5. Sharma, A.K., Ling, C.Y., Birgersson, E., Vynnycky, M., Han, M.: Verified reduction of dimensionality for an all-vanadium redox flow battery model. J. Power Sources **279**, 345–350 (2015)
6. Vynnycky, M., Assunção, M.: The vanadium redox flow battery: an asymptotic perspective. SIAM J. Appl. Math. **79**, 1147–1172 (2019)
7. Vynnycky, M., Assunção, M.: On the significance of sulphuric-acid dissociation in the modelling of vanadium redox flow batteries. J. Engng. Math. **123**, 173–203 (2020)
8. Assunção, M.: Mathematical modelling of vanadium redox batteries. Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden (2015)
9. Knehr, K.W., Agar, E., Dennison. C.R., Kalidindi. A.R., Kumbur, E.C.: A transient vanadium flow battery model incorporating vanadium crossover and water transport through the membrane. J. Electrochem. Soc. **159**, A1446–A1459 (2012)
10. You, D., Zhang, H., Chen, J.: A simple model for the vanadium redox battery. Electrochim. Acta **54**, 6827–6836 (2009)
11. Ma, X., Zhang, H., Xing, F.: A three-dimensional model for negative half cell of the vanadium redox flow battery. Electrochim. Acta **58**, 238–246 (2011)

Time-Adaptive High-Order Compact Finite Difference Schemes for Option Pricing in a Family of Stochastic Volatility Models

Bertram Düring and Christof Heuer

Abstract We propose a *time-adaptive high-order compact finite difference scheme* for option pricing in a *family of stochastic volatility models*. We employ a semi-discrete high-order compact finite difference method for the spatial discretisation, and combine this with an adaptive time discretisation, extending ideas from Lötstedt et al. (Implicit solution of hyperbolic equations with space-time adaptivity, BIT, 42(1):134–158, 2002.) to fourth-order multistep methods in time.

1 Introduction

Stochastic volatility models have become one of the standard approaches for financial option pricing. They are based on a two-dimensional stochastic diffusion process containing two Brownian motions with correlation $\rho \in [-1, 1]$, i.e. $\mathrm{E}\big[\mathrm{d}W_1(t)\mathrm{d}W_2(t)\big] = \rho\, dt$, on a given filtered probability space for the underlying asset $S = S(t)$ and the stochastic variance $v = v(t)$. In this work we consider the following class of stochastic volatility models,

$$dS = \mu S\,\mathrm{d}t + \sqrt{v}S\,\mathrm{d}W_1, \quad dv = \kappa v^a\,(\theta - v)\,\mathrm{d}t + \sigma v^b \mathrm{d}W_2, \tag{1}$$

with given drift $\mu \in \mathbb{R}$ of the underlying $S(t)$, long run mean $\theta > 0$, mean reversion speed $\kappa > 0$, and volatility of volatility $\sigma > 0$, see e.g. [1]. Additionally, it holds $a \geq 0$ and $b \in (0, 3/2]$. Many well-known models are included in the family (1). The prominent *Heston (or SQR) model* [5] is obtained for $a = 0$, $b = 1/2$. Other known models include the *GARCH (or VAR) model* [4], with $a = 0$, $b = 1$, and

B. Düring (✉)
Mathematics Institute, University of Warwick, Coventry, UK
e-mail: bertram.during@warwick.ac.uk

C. Heuer
d-fine GmbH, Frankfurt, Germany
e-mail: heuer.chr@googlemail.com

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_49

the *3/2-model* [7] in which $a = 0$, $b = 3/2$. There are also models with *non-linear* mean reversion, following [1], we denote these models as the *SQR-N model* ($a = 1$, $b = 1/2$), *VAR-N model* ($a = 1$, $b = 1$), and 3/2-*N model* ($a = 1$, $b = 3/2$).

For the family of stochastic volatility models (1), application of Itô's Lemma and standard arbitrage arguments lead to partial differential equations for the option price $V = V(S, v, t)$, which are of the following form

$$\frac{\partial V}{\partial t} + \frac{vS^2}{2}\frac{\partial^2 V}{\partial S^2} + \rho\sigma v^{b+\frac{1}{2}} S \frac{\partial^2 V}{\partial S \partial v} + \frac{\sigma^2 v^{2b}}{2}\frac{\partial^2 V}{\partial v^2} + rS\frac{\partial V}{\partial S} + \kappa v^a (\theta - v)\frac{\partial V}{\partial v} - rV = 0, \qquad (2)$$

where $r \geq 0$ denotes the risk-free interest rate. Equation (2) has to be solved (backward in time) for $S, v > 0$, $0 \leq t < T$, with an expiration date $T > 0$, and subject to final and boundary conditions depending on the specific option considered. In the case of a European Put options, for example, the final condition is given by $V(S, v, T) = \max(K - S, 0)$ with strike price $K > 0$.

In the mathematical literature, there are many works on numerical methods for option pricing in one-dimension (single risk factor), but less papers considering numerical methods for option pricing in stochastic volatility models, i.e. for two spatial dimensions. Finite difference approaches used are often standard, low order methods, i.e. second order in space. In the last decade, high-order (fourth order in space) compact finite difference discretisations for option pricing in stochastic volatility models have been presented, e.g. in [2, 3]. We refer to [3] for an overview of the finite difference literature and other methods.

The originality of the present chapter consists in proposing a new, *time-adaptive high-order compact finite difference scheme* for option pricing in a *family of stochastic volatility models*. Our approach builds on ideas from [3] and [8]. We employ a semi-discrete high-order compact finite difference method for the spatial discretisation, using the methodology developed in [3]. For the adaptive time discretisation, we follow basic ideas of [8], where two-step methods for the time-discretisation were used, and generalise this approach to consider fourth-order multistep methods in time. We obtain a time-adaptive high-order compact scheme that is fourth order accurate in both space and time.

2 Transformation of the Partial Differential Equation

We first transform $\tau = T - t$, and $u = \exp(r\tau)V/K$ in (2). Depending on the model parameter b, we apply subsequent transformations, in such a way that the second derivatives in x- and y-direction share the same coefficient.

For $b \neq 3/2$ we apply the transformations $x = (3/2 - b)\ln(S/K)$, $y = v^{3/2-b}/\sigma$, and arrive at

$$u_\tau + c_1(y)\left(u_{xx} + u_{yy}\right) + c_2(y)u_{xy} + c_3(y)u_x + c_4(y)u_y = 0, \tag{3}$$

to be solved on the rectangular spatial domain $\Omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$, with

$$\begin{aligned}
c_1(y) &= -\,\sigma^{\frac{-5+2b}{-3+2b}}\, y^{-2\,(-3+2\,b)^{-1}}\,(-3+2\,b)^2\,(8\sigma)^{-1}, \quad c_2(y) = 2\rho c_1(y),\\
c_3(y) &=(3-2b)\big(\sigma^{\frac{-5+2b}{-3+2b}}\, y^{-2\,(-3+2\,b)^{-1}} - 2\,r\sigma\big)(4\sigma)^{-1},\\
c_4(y) &=(3-2b)\Big(2\,\sigma^{\frac{-5+2b}{-3+2b}}\, y^{-\frac{-1+2b}{-3+2b}}\, b - 4\,\sigma^{-\frac{1+2a-2b}{-3+2b}}\, y^{-\frac{1+2a-2b}{-3+2b}}\,\kappa\,\theta\\
&\qquad + 4\,\sigma^{-\frac{3+2a-2b}{-3+2b}}\, y^{-\frac{3+2a-2b}{-3+2b}}\,\kappa - \sigma^{\frac{-5+2b}{-3+2b}}\, y^{-\frac{-1+2b}{-3+2b}}\Big)(8\sigma)^{-1},
\end{aligned}$$

and subject to $u(x, y, 0) = \max\left(1 - \exp\left(x/(3/2 - b)\right), 0\right).$

For $b = 3/2$, we apply the transformations $x = \ln(S/K)$, $y = \ln(v)/\sigma$, and obtain (3) with coefficients $c_1(y) = -\exp(\sigma y)/2$, $c_2(y) = -\rho\exp(\sigma y)$, $c_3(y) = \exp(\sigma y)/2 - r$, $c_4(y) = (\sigma^2\exp(\sigma y) - 2\kappa\theta\exp(\sigma y(a-1)) + 2\kappa\exp(a\sigma y))/(2\sigma)$ and subject to $u(x, y, 0) = \max\left(1 - \exp(x), 0\right).$

3 Time-Adaptive High-Order Compact Scheme

We use the high-order compact semi-discrete (discretising in space only) scheme from [3] for (3). Since the coefficients of u_{xx} and u_{yy} in (3) are identical, results from [3] show that the scheme provides a fourth-order accurate spatial discretisation employing a uniform grid with $h_1 = h_2 = h > 0$. The semi-discrete scheme can be written in matrix form as

$$M_h \partial_\tau U_h(\tau) = g^{(h)}(\tau) - K_h U^{(h)}(\tau) =: F(\tau). \tag{4}$$

The known vector $g^{(h)}$ has only non-zero entries due to the influence of the boundary conditions and the matrices M_h and K_h do not depend on τ.

At the boundary $x = x_{\min}$ and $x = x_{\max}$ we impose Dirichlet type boundary conditions. For $y = y_{\min}$ or $y = y_{\max}$ we do not impose any boundary condition, but apply the discretisation of the spatial interior. The resulting ghost points are extrapolated from the interior with sufficiently high order. Due to the low regularity of the typical initial conditions, we employ a smoothing operator [6] to ensure fourth-order spatial convergence. For further details of the implementation of boundary and initial conditions, we refer to [3].

Our approach for time adaptivity is motivated by [8], where two-step methods are used for time discretisation. Here, to match the fourth-order accuracy in space, we consider fourth-order multistep methods in time. We approximate the system of ordinary differential equations (4) using fourth-order multistep methods and variable, adaptive time step sizes. In each time step, we use a (numerically cheap) *predictor* scheme to estimate the local truncation error, adapt the time step accordingly, and then solve using a *corrector* scheme. Necessary start-up values are computed using a Crank-Nicolson time-discretisation.

Predictor Scheme Consider $\tau_{\min} = \tau_0 < \tau_1 < \ldots < \tau_j$ with $j \geq 4$ and $\tau_j < \tau_{\max}$ in time with the step sizes $k_n = \tau_n - \tau_{n-1} > 0$ for $n = 1, \ldots, j$. We denote the value of the vector $U^{(h)}$ at time τ_n by $U_n^{(h)}$.

We use a four-step predictor scheme with (non-equidistant) time steps,

$$\alpha_0^{(\text{pre})} M_h U_n^{(h)} = k_n g_{n-1}^{(h)} - \left[\alpha_1^{(\text{pre})} M_h + k_n K_h\right] U_{n-1}^{(h)} - M_h \sum_{j=2}^{4} \alpha_j^{(\text{pre})} U_{n-j}^{(h)}, \tag{5}$$

where

$$\alpha_0^{(\text{pre})} = 2\,\iota_1\iota_2\iota_3 + \iota_3\iota_1^2 + \iota_2\iota_1^2 + \iota_2^2\iota_3 + \iota_1\iota_2^2/\varphi_0^{(\text{pre})},$$

$$\alpha_1^{(\text{pre})} = \frac{\iota_1^3\iota_2 - 2\,\iota_1\iota_2\iota_3 - \iota_3\iota_1^2 - \iota_2\iota_1^2 - \iota_2^2\iota_3 - \iota_1\iota_2^2 + 3\,\iota_2^2\iota_3\iota_1 + 4\,\iota_1^2\iota_3\iota_2 + 2\,\iota_1^2\iota_2^2 + \iota_1^3\iota_3}{2\,\iota_1\iota_2\iota_3 + \iota_3\iota_1^2 + \iota_2\iota_1^2 + \iota_2^2\iota_3 + \iota_1\iota_2^2},$$

$$\alpha_2^{(\text{pre})} = -\,2\,\iota_1\iota_2\iota_3 + \iota_3\iota_1^2 + \iota_2\iota_1^2 + \iota_2^2\iota_3 + \iota_1\iota_2^2/((\iota_2 + \iota_3)\,(\iota_1 + 1)),$$

$$\alpha_3^{(\text{pre})} = \iota_2^2\,(\iota_1\iota_2 + \iota_1\iota_3 + \iota_2\iota_3)/((\iota_1\iota_2 + \iota_2 + \iota_1)\,(\iota_2 + \iota_1)),$$

$$\alpha_4^{(\text{pre})} = -\,(\iota_2 + \iota_1)\,\iota_2^2\iota_3^4/((\iota_1\iota_2\iota_3 + \iota_2\iota_3 + \iota_1\iota_3 + \iota_1\iota_2)\,(\iota_1\iota_2 + \iota_1\iota_3 + \iota_2\iota_3)\,(\iota_2 + \iota_3)),$$

with $\iota_1 = k_n/k_{n-1}$, $\iota_2 = k_n/k_{n-2}$, $\iota_3 = k_n/k_{n-3}$, as well as

$$\begin{aligned}\varphi_0^{(\text{pre})} = {} & \iota_1^3\iota_3\iota_2^2 + 3\,\iota_2^2\iota_3\iota_1 + 4\,\iota_1^2\iota_3\iota_2 + 2\,\iota_1^3\iota_3\iota_2 + 3\,\iota_1^2\iota_2^2\iota_3 + \iota_2^2\iota_1^3 + 2\,\iota_1^2\iota_2^2 \\ & + \iota_2^2\iota_3 + 2\,\iota_1\iota_2\iota_3 + \iota_1\iota_2^2 + \iota_3\iota_1^2 + \iota_1^3\iota_3 + \iota_1^3\iota_2 + \iota_2\iota_1^2.\end{aligned}$$

The predictor scheme (5) is implicit. However, since M_h does not depend on τ, it has to be factorised only once at the beginning and the factorisation can then be reused in every time step. Hence, the predictor scheme is still computationally cheap.

The local truncation error of the predictor scheme is given by

$$U^{(h)}(\tau_n) - \tilde{U}_n^{(h)} = C_P^{\text{loc}} k_n^5 \frac{\partial^5 u}{\partial \tau^5} + \mathcal{O}\left(k_n^6\right), \tag{6}$$

with $C_P^{\text{loc}} = [(\iota_1 + 1)\,(\iota_1\iota_2 + \iota_2 + \iota_1)\,(\iota_1\iota_2\iota_3 + \iota_2\iota_3 + \iota_1\iota_3 + \iota_1\iota_2)]/[120{\iota_1}^3\iota_3{\iota_2}^2]$.

In the following, we use the notation $\tilde{U}_n^{(h)}$ to clarify whenever the predictor scheme is used to obtain the approximation of the solution $U^{(h)}(\tau_n)$.

Corrector Scheme For the corrector step, we use the implicit BDF-4 method with variable step-sizes to approximate the system of ordinary differential equations (4),

$$\left[\alpha_0^{(\mathrm{cor})} M_h + k_n K_h\right] U_n^{(h)} = - M_h \sum_{j=1}^{4} \alpha_j^{(\mathrm{cor})} U_{n-j}^{(h)} + k_n g_n^{(h)}, \tag{7}$$

where

$$\begin{aligned}
\alpha_0^{(\mathrm{cor})} =& \frac{3\,\iota_2^2\iota_1^3 + 4\,\iota_1^3\iota_3\iota_2^2 + 6\,\iota_1^3\iota_3\iota_2 + 2\,\iota_1^3\iota_2 + 2\,\iota_1^3\iota_3 + 9\,\iota_1^2\iota_2^2\iota_3 + 4\,\iota_1^2\iota_2^2 + \iota_2\iota_1^2}{(\iota_1\iota_2\iota_3 + \iota_2\iota_3 + \iota_1\iota_3 + \iota_1\iota_2)\,(\iota_1\iota_2 + \iota_2 + \iota_1)\,(\iota_1 + 1)} \\
&+ \frac{8\,\iota_1^2\iota_3\iota_2 + \iota_3\iota_1^2 + 6\,\iota_2^2\iota_3\iota_1 + \iota_1\iota_2^2 + 2\,\iota_1\iota_2\iota_3 + \iota_2^2\iota_3}{(\iota_1\iota_2\iota_3 + \iota_2\iota_3 + \iota_1\iota_3 + \iota_1\iota_2)\,(\iota_1\iota_2 + \iota_2 + \iota_1)\,(\iota_1 + 1)}, \\
\alpha_1^{(\mathrm{cor})} =& - \frac{3\,\iota_2^2\iota_3\iota_1 + 4\,\iota_1^2\iota_3\iota_2 + 2\,\iota_1^3\iota_3\iota_2 + 3\,\iota_1^2\iota_2^2\iota_3 + \iota_1^3\iota_3\iota_2^2 + \iota_2^2\iota_1^3 + 2\,\iota_1^2\iota_2^2 + \iota_2^2\iota_3}{(\iota_1\iota_2 + \iota_1\iota_3 + \iota_2\iota_3)\,(\iota_2 + \iota_1)} \\
&- \frac{2\,\iota_1\iota_2\iota_3 + \iota_1\iota_2^2 + \iota_3\iota_1^2 + \iota_1^3\iota_3 + \iota_1^3\iota_2 + \iota_2\iota_1^2}{(\iota_1\iota_2 + \iota_1\iota_3 + \iota_2\iota_3)\,(\iota_2 + \iota_1)}, \\
\alpha_2^{(\mathrm{cor})} =& \frac{\iota_1^2\iota_2^2 + \iota_1^2\iota_2^2\iota_3 + 2\,\iota_1^2\iota_3\iota_2 + \iota_2\iota_1^2 + \iota_3\iota_1^2 + 2\,\iota_2^2\iota_3\iota_1 + \iota_1\iota_2^2 + 2\,\iota_1\iota_2\iota_3 + \iota_2^2\iota_3}{(\iota_1 + 1)\,(\iota_2 + \iota_3)}, \\
\alpha_3^{(\mathrm{cor})} =& - \frac{\left(\iota_2\iota_3 + \iota_1\iota_3 + \iota_3\iota_1^2 + 2\,\iota_1\iota_2\iota_3 + \iota_1^2\iota_3\iota_2 + \iota_2\iota_1^2 + \iota_1\iota_2\right)\iota_2^2}{\iota_2\iota_1^2 + \iota_1\iota_2^2 + 2\,\iota_1\iota_2 + \iota_2^2 + \iota_1^2}, \\
\alpha_4^{(\mathrm{cor})} =& \left(\iota_2 + \iota_1 + \iota_1^2 + 2\,\iota_1\iota_2 + \iota_2\iota_1^2\right)\iota_2^2\iota_3^4/\varphi_4^{(\mathrm{cor})},
\end{aligned}$$

with $\iota_1 = k_n/k_{n-1}$, $\iota_2 = k_n/k_{n-2}$, $\iota_3 = k_n/k_{n-3}$, as well as

$$\begin{aligned}
\varphi_4^{(\mathrm{cor})} = &\ \iota_3^3\iota_1^2 + 2\,\iota_1\iota_3^3\iota_2 + 4\,\iota_1\iota_2^2\iota_3^2 + \iota_2^2\iota_3^3 + 2\,\iota_1^2\iota_2^2\iota_3^2 + \iota_2^2\iota_3^3\iota_1 + \iota_2^3\iota_3^2 \\
&+ \iota_1^2\iota_2^3 + \iota_2\iota_3^3\iota_1^2 + \iota_2^3\iota_3^2\iota_1 + 3\,\iota_1^2\iota_2\iota_3^2 + 3\,\iota_1^2\iota_2^2\iota_3 + \iota_2^3\iota_3\iota_1^2 + 2\,\iota_1\iota_2^3\iota_3.
\end{aligned}$$

The local truncation error of the corrector scheme is given by

$$U^{(h)}\left(\tau_n\right) - U_n^{(h)} = C_C^{\mathrm{loc}} k_n^5 \frac{\partial^5 U^{(h)}(\tau_n)}{\partial \tau^5} + O\left(k_n^6\right), \tag{8}$$

with

$$\begin{aligned}
C_C^{\mathrm{loc}} =& - \left(\iota_1\iota_2\iota_3 + \iota_2\iota_3 + \iota_1\iota_3 + \iota_1\iota_2\right)^2 \left(\iota_1 + 1\right)^2 \left(\iota_1\iota_2 + \iota_2 + \iota_1\right)^2/N_C^{\mathrm{loc}}, \\
N_C^{\mathrm{loc}} =& 120\iota_1^3\iota_2^2\iota_3 \Big[4\,\iota_1^3\iota_3\iota_2^2 + 6\,\iota_3\iota_1^3\iota_2 + 2\,\iota_3\iota_1^3 + 3\,\iota_2^2\iota_1^3 + 2\,\iota_2\iota_1^3 + 8\,\iota_1^2\iota_3\iota_2 \\
&+9\,\iota_1^2\iota_2^2\iota_3 + \iota_1^2\iota_3 + 4\,\iota_1^2\iota_2^2 + \iota_2\iota_1^2 + 6\,\iota_2^2\iota_3\iota_1 + 2\,\iota_1\iota_2\iota_3 + \iota_1\iota_2^2 + \iota_3\iota_2^2\Big].
\end{aligned}$$

Time-Step Adaption The aim of the time-step adaption is to choose the time-step in such a way that the resulting local time-discretisation error stays below a given threshold $\hat{\epsilon} > 0$. Similar as in [9], we use the local time-discretisation errors (6) and (8) to obtain the first order approximation

$$\frac{\partial^5 U^{(h)}(\tau_n)}{\partial \tau^5} = \frac{U_n^h - \tilde{U}_n^h}{k_n^5 \left(C_C^{\text{loc}} - C_P^{\text{loc}}\right)} + O(k_n). \tag{9}$$

The leading error term of the discretisation (7) can thus be approximated by

$$\epsilon_n = -\alpha_0^{(\text{cor})} M_h C_C^{(\text{loc})} k_n^4 \frac{\partial^5 U^{(h)}}{\partial \tau^5} = -\alpha_0^{(\text{cor})} M_h C_C^{(\text{loc})} \frac{U_n^h - \tilde{U}_n^h}{k_n \left(C_C^{\text{loc}} - C_P^{\text{loc}}\right)}. \tag{10}$$

The goal is now to choose the next step-size in time in a way that the norm of this error is bounded by the error threshold $\hat{\epsilon} > 0$ in a given norm. The general error structure is given by $\epsilon_n = k_n^4 \zeta(\tau_n) \iff k_n = (\epsilon_n / \zeta(\tau_n))^{\frac{1}{4}}$ (with $\zeta(\tau_n)$ implicitly defined by (10)) and thus we can, with $\|\epsilon_n\| \leq \hat{\epsilon}$, use $k_{n+1} \leq k_n(\hat{\epsilon}/\|\epsilon_n\|)^{\frac{1}{4}}$ to choose the new step size in time.

The approximation of the local discretisation error in time (10) can be non-smooth, giving rise to abrupt changes of the chosen step size. To ensure that we avoid choosing a very large step size in case that the estimated error is very small, we introduce a small parameter $\beta > 0$ (see [9]) and adapt the time step size according to

$$k_{n+1} = \left(\frac{\hat{\epsilon}}{\hat{\epsilon}\beta + \|\epsilon_n\|}\right)^{\frac{1}{4}} k_n =: \xi_n k_n. \tag{11}$$

4 Numerical Results

We consider the pricing of European Put options with model (1) and use $(S, v) \in (1.5, 600) \times (0.1, 0.5)$. The computational domain is determined through the transformations given in Sect. 2. We choose step-size $h = (x_{\max} - x_{\min})/(N-1)$ with $N = 201$ steps in x-direction, in y-direction we begin at $y_{\min}$ and use step-size h. In (11), we set $\beta = 0.01$. We use $K = 100$, $T = 2$, $r = 0.05$, $\sigma = 0.3$, $\kappa = 1.1$, $\theta = 0.3$, $\rho = -0.4$. For the start-up values, we apply the Crank-Nicolson time-steps with a fixed parabolic mesh ratio, choosing $k_n = 0.05h^2$, $n = 1, 2, 3$.

Figure 1 shows the adaptation factor ξ_n, the positioning of the grid points in time, and the local error $||\epsilon_n||_2$ for the GARCH model (left column) and the $a = b = 3/4$ model (right column). For GARCH the algorithm leads to overall 104 grid-points in

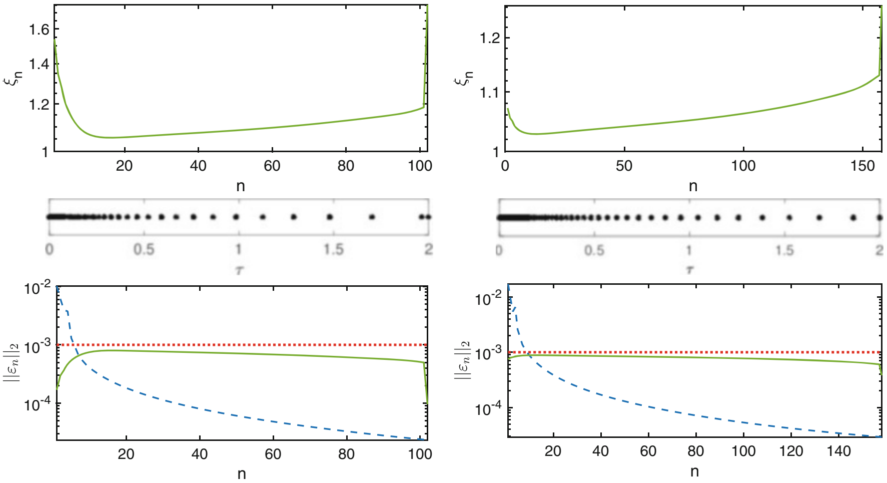


Fig. 1 Adaptation factor ξ_n, time grid points distribution, and error threshold $\hat{\epsilon}$ (dotted red), local error $||\epsilon_n||_2$ for adaptive (solid green) and equidistant time stepping (dashed blue): GARCH (left), $a = b = 3/4$ model (right)

time. The local error remains just below the chosen threshold $\hat{\epsilon} = 0.001$, while time steps are increased. For GARCH, 50 of 104 grid-points in time, including the three initial points where Crank-Nicolson type time discretisation is used, are located in the interval $[0, 0.01]$, i.e. 48% of the grid-points are positioned in only 0.5% of the time-domain. On the other hand only six points are placed in the time interval $[1, 2]$. The results for the $a = b = 3/4$ model show a similar behaviour. For comparison we repeat both simulations, now with the same numbers of *equidistant* time steps. Initially, the local error is above the threshold and later far below, indicating the sub-optimality of the equidistant distribution of points in time.

References

1. P. Christoffersen, K. Jacobs, and K. Mimouni, *Models for S&P500 dynamics: Evidence from realized volatility, daily returns, and option prices*, Review of Financial Studies, 23:3141–3189, 2010.
2. B. Düring and M. Fournié, *High-order compact finite difference scheme for option pricing in stochastic volatility models*, J. Comput. Appl. Math., 236(17):4462–4473, 2012.
3. B. Düring and C. Heuer, *High-order compact schemes for parabolic problems with mixed derivatives in multiple space dimensions*, SIAM J. Numer. Anal., 53(5):2113–2134, 2015.
4. J. Duan, *The GARCH option pricing model*, Math. Finance, 5(1):13–32, 1995.
5. S.L. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Rev. Fin. Studies, 6(2):327–343, 1993.
6. H.O. Kreiss, V. Thomee, and O. Widlund, *Smoothing of initial data and rates of convergence for parabolic difference equations*, Commun. Pure Appl. Math., 23:241–259, 1970.
7. A.L. Lewis, *Option valuation under stochastic volatility*, Finance Press, Newport Beach, CA, 2000.

8. P. Lötstedt, S. Söderberg, A. Ramage, and L. Hemmingsson-Frändén, *Implicit solution of hyperbolic equations with space-time adaptivity*, BIT, 42(1):134–158, 2002.
9. J. Persson and L. von Sydow, *Pricing European multi-asset options using a space-time adaptive FD-method*, Computing and Visualization in Science, 10:173–183, 2007.

Multirate DAE-Simulation and Its Application in System Simulation Software for the Development of HVAC Systems

Michael Kolmbauer, Günter Offner, and Bernhard Pöchtrager

Abstract This work is devoted to the efficient simulation of large multi-physical networks stemming from automated modeling processes in system simulation software. The simulation of heating, ventilation and air conditioning (HVAC) applications for passenger cars requires the coupling of gas, fluid and thermal networks. Each network is established by combining the connection structure of a graph with physical equations of elementary components and resulting in a differential algebraic equation (DAE). In order to speed up the simulation, a non-iterative multirate time integration co-simulation method for the system of coupled DAEs is introduced. The power of the multirate method is shown via a representative example of a HVAC vehicle cabin model, which simulates the cooling and heating of the air flow and its circulation in and out of the passenger compartment.

1 Problem Formulation

We consider a network that is composed of multi-physical components. The network elements describing the gas contribution are given by resistive elements, compressors, nodes, system boundaries, mass flow terminations, heat transfers and temperature boundaries. The fluid network consists of pipes, pumps, demands, junctions and reservoirs. The thermal coupling is established by lumped mass

M. Kolmbauer
MathConsult GmbH, Linz, Austria
e-mail: michael.kolmbauer@mathconsult.co.at

G. Offner
AVL List GmbH, Graz, Austria
e-mail: guenter.offner@avl.com

B. Pöchtrager (✉)
Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria
e-mail: bernhard.poechtrager@ricam.oeaw.ac.at

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_50

elements representing the pipe wall and the masses from the heat exchangers and heat transfer connections. The individual components are assembled to a network $\mathcal{N}$, which is represented by a linear directed graph. The graph structure is described by an incidence matrix A, which can be used for the model descriptions. In the following we state the DAEs for the three main involved physical networks.

Fluid Network
We consider a fluid network $\mathcal{N}_F = \{PI, PU, DE, VJ, LJ, RE, HT_F, TB_F\}$ that is composed of pipes PI, pumps PU, demands DE, volume junctions VJ, lumped junctions LJ, reservoirs RE, heat transfers HT_F and temperature boundaries TB_F. The DAE for the network $\mathcal{N}_F$ in input-output form is given by: For given continuous inputs $(u_{Hs_F}^T, u_{Tb_F}^T)^T$, find the pressures $(p_{Lj}^T, p_{Vj}^T)^T$ the mass flows $(q_{Pi}^T, q_{Pu}^T)^T$, the temperatures $(T_{Vj}^T, T_{Lj}^T)^T$, the heat fluxes $(H_{Ht_F}^T, H_{Pu}^T, H_{Pi}^T)^T$ and the outputs $(y_{Vj}^T, y_{Lj}^T, y_{Ht_F}^T)^T$, such that

$$\begin{aligned}
\frac{dq_{Pi}}{dt} &= c_{1,Pi}\left(A_{Jc,Pi}^T p_{Jc} + A_{Re,Pi}^T p_{Re}\right) + c_{2,Pi}\operatorname{diag}\left(|q_{Pi}|\right) q_{Pi} + c_{3,Pi} \\
f_{Pu}(q_{Pu}) &= A_{Jc,Pu}^T p_{Jc} + A_{Re,Pu}^T p_{Re} \\
0 &= A_{Jc,Pi} q_{Pi} + A_{Jc,Pu} q_{Pu} + A_{Jc,De} q_{De} \\
m_{Vj} c_{p,Vj} \frac{dT_{Vj}}{dt} &= A_{Vj,Pi} H_{Pi} + A_{Vj,Pu} H_{Pu} \\
&\quad + A_{Vj,De} H_{De} + A_{Vj,Ht_F} H_{Ht_F} + A_{Vj,Hs_u} u_{Hs_F} \\
0 &= A_{Lj,Pi} H_{Pi} + A_{Lj,Pu} H_{Pu} \\
&\quad + A_{Lj,De} H_{De} + A_{Lj,Ht_F} H_{Ht_F} + A_{Lj,Hs_u} u_{Hs_F} \\
H_{Pi} &= B_{Jc}(q_{Pi}) T_{Vj} + B_{Jc}(q_{Pi}) T_{Lj} + B_{Jc}(q_{Pi}) T_{Re} \\
H_{Pu} &= B_{Jc}(q_{Pu}) T_{Vj} + B_{Jc}(q_{Pu}) T_{Lj} + B_{Jc}(q_{Pu}) T_{Re} \\
H_{Ht_F} &= c_{Ht_F}\left(A_{Vj,Ht_F}^T T_{Vj} + A_{Lj,Ht_F}^T T_{Lj} + A_{Tb_u,Ht_F}^T u_{Tb_F}\right) \\
y_{Vj} &= |(A_{Vj,Hs_F}^T + A_{Tb_u,Ht_F} A_{Vj,Ht_F}^T)| T_{Vj} \\
y_{Lj} &= |(A_{Lj,Hs_F}^T + A_{Tb_u,Ht_F} A_{Lj,Ht_F}^T)| T_{Lj} \\
y_{Ht_F} &= (A_{Tb_F,Ht_F} + A_{Lj,Hs_F}^T A_{Lj,Ht_F}^T + A_{Vj,Hs_F}^T A_{Vj,Ht_F}^T) H_{Ht_F}
\end{aligned} \tag{1}$$

for given boundary conditions $q_{De} = \bar{q}_{De}$, $H_{De} = \bar{H}_{De}$, $p_{Re} = \bar{p}_{Re}$ and $T_{Re} = \bar{T}_{Re}$ and given coefficients $c_{1,Pi}$, $c_{2,Pi}$, $c_{3,Pi}$, m_{Vj}, $c_{p,Vj}$ and c_{Ht_F} as well as given functions f_{Pu}. The function B_{Jc} checks for the sign of the mass flow q_{Pi}. The coupling variables are given by the temperatures u_{Hs_F} and u_{Tb_F} and the energy fluxes y_{Vj}, y_{Lj} and y_{Ht_F}.

Gas Network

We consider a gas network $\mathcal{N}_G = \{R, C, N, SB, MT, HT_G, TB_G\}$ that is composed of resistive elements R, compressors C, nodes N, system boundaries SB, mass flow terminations MT, heat transfers HT_G and temperature boundaries TB_G. The DAE for the network $\mathcal{N}_G$ in input-output form is given by: For given continuous inputs $(u_{Hs_G}^T, u_{Tb_G}^T)^T$, find the pressures $(p_N^T)^T$ the mass flows $(q_R^T, q_C^T)^T$, the temperatures $(T_N^T)^T$, the heat fluxes $(H_{Ht_G}^T, H_C^T, H_R^T)^T$ and the outputs $(y_N^T, y_{Ht_G}^T)^T$, such that

$$
\begin{aligned}
\frac{dq_R}{dt} &= c_{1,R}\left(A_{N,R}^T p_N + A_{Sb,R}^T p_{Sb}\right) + c_{2,R}\text{diag}\frac{(|q_R|)\, q_R}{p_{up}(p_N, p_{Sb})} + c_{3,R} \\
f_C(q_C) &= A_{N,C}^T p_N + A_{Sb,C}^T p_{Sb} \\
0 &= A_{N,R} q_R + A_{N,C} q_C + A_{N,Mt} q_{Mt} \\
0 &= A_{N,R} H_R + A_{N,C} H_C + A_{N,Mt} H_{Mt} + A_{N,Ht_G} H_{Ht_G} + A_{N,Hs_u} u_{Hs_G} \\
H_R &= B_N(q_R) T_N + B_N(q_R) T_{Sb} \\
H_C &= B_N(q_C) T_N + B_N(q_C) T_{Sb} \\
H_{HT_G} &= c_{Ht_G}\left(A_{N,Ht_G}^T T_N + A_{Sb_u,Ht_G}^T u_{Sb_G}\right) \\
y_N &= |(A_{N,Hs_G}^T + A_{Tb_G,Ht_G} A_{N,Ht_G}^T)| T_N \\
y_{Ht_G} &= (A_{Tb_G,Ht_G} + A_{N,Hs_G}^T A_{N,Ht_G}^T) H_{Ht_G}
\end{aligned}
\tag{2}
$$

for given boundary conditions $q_{Mt} = \bar{q}_{Mt}$, $H_{Mt} = \bar{H}_{Mt}$, $p_{Sb} = \bar{p}_{Sb}$ and $T_{Sb} = \bar{T}_{Sb}$, and given coefficients $c_{1,R}$, $c_{2,R}$, $c_{3,R}$ and c_{Ht_G} as well as given functions f_C, B_N and p_{up}. The coupling variables are expressed as the temperatures u_{Hs_G} and u_{Tb_G} and the energy fluxes y_N and y_{Ht_G}.

Solid Network

We consider a solid network $\mathcal{N}_S = \{SW, LW, HT_S, HS, TB_S\}$ that is composed of solid walls SW, lumped walls LW, heat transfers HT_S, heat sources HS and temperature boundaries TB_S. The DAE for the network $\mathcal{N}_S$ in input-output form is given by: For given continuous inputs $(u_{Hs_S}^T, u_{Tb_S}^T)^T$, find the temperatures $(T_{Sw}^T, T_{Lw}^T)^T$, the heat fluxes $(H_{Ht_S}^T)^T$ and the outputs $(y_{Sw}^T, y_{Lw}^T, y_{Ht_S}^T)^T$, such that

$$\begin{aligned}
m_{Sw} c_{p,Sw} \frac{dT_{Sw}}{dt} &= A_{Sw,Ht_S} H_{Ht_S} + A_{Sw,Hs} H_{Hs} + A_{Sw,Hs_u} u_{Hs_S} \\
0 &= A_{Lw,Ht_S} H_{Ht_S} + A_{Lw,Hs} H_{Hs} + A_{Lw,Hs_u} u_{Hs_S} \\
H_{Ht_S} &= c_{Ht_S} \left(A^T_{Sw,Ht_S} T_{Sw} + A^T_{Lw,Ht_S} T_{Lw} + A^T_{Tb,Ht_S} T_{Tb} + A^T_{Tb_u,Ht_S} u_{Tb_S} \right) \\
y_{Sw} &= |(A^T_{Sw,Hs_S} + A_{Tb_S,Ht_S} A^T_{Sw,Ht_S})| T_{Sw} \\
y_{Lw} &= |(A^T_{Lw,Hs_S} + A_{Tb_S,Ht_S} A^T_{Lw,Ht_S})| T_{Lw} \\
y_{Ht_S} &= A_{Tb_S,Ht_S} H_{Ht_S}
\end{aligned} \tag{3}$$

for given boundary conditions $H_{Hs} = \bar{H}_{Hs}$ and $T_{Tb} = \bar{T}_{Tb}$ and given positive definite coefficient matrices m_{Sw}, $c_{p,Sw}$ and c_{Ht_S}. The coupling variables are expressed as the energy fluxes u_{Hs_S} and u_{Tb_S} and the temperatures y_{Sw}, y_{Lw} and y_{Ht_S}.

Multi-Physical Model

The multi-physical model is derived by combining (1), (2) and (3) with appropriate coupling conditions. The coupling conditions describe the relation between the inputs and outputs of the individual models. For the model used in Sect. 3, the following coupling conditions are used, see e.g. [4].

$$\begin{pmatrix} u_{Hs_G} \\ u_{Tb_G} \\ u_{Hs_S} \\ u_{Tb_S} \\ u_{Hs_F} \\ u_{Tb_F} \end{pmatrix} = \begin{pmatrix} 0 & C_{Hs_G,N} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_{Tb_G,Sw} & 0 & 0 & 0 & 0 & 0 \\ C_{Hs_S,N} & 0 & 0 & 0 & 0 & 0 & 0 & C_{Hs_S,Ht_F} \\ 0 & 0 & 0 & 0 & 0 & C_{Tb_S,Vj} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{Hs_F,Vj} & 0 & 0 & 0 \\ 0 & 0 & C_{Tb_F,Sw} & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_N \\ y_{Ht_G} \\ y_{Sw} \\ y_{Lw} \\ y_{Ht_S} \\ y_{Vj} \\ y_{Lj} \\ y_{Ht_F} \end{pmatrix} \tag{4}$$

The connectivity equation (4) represents the thermal coupling of the fluid and gas network. Combining all subsystems and their connectivity equations (4) yields the resulting DAE.

DAEs resulting from automated modeling software typically obtain a structure with (differential) index greater 1, cf. [2, 3] and hence are not suitable for a direct simulation with standard solvers. In the setup of multiple physical networks it is not sufficient, that the full DAE can be reduced to a d-index (differential index) 1. Additionally, each subsystem, to which a solver is applied, has to fulfill d-index 1 conditions as well, cf. [1]. In our applications an automatic index reduction is performed if the gas or the fluid system happens to be of d-index 2.

2 Multirate Integration for Coupled Network DAEs

Based on the physical background the full DAE is partitioned to $n \in \mathbb{N}$ subsystems (typically $n \gg 2$) in our multirate approach. Each subsystem is index reduced according to the available literature, cf. [2, 3]. Since in the global network the individual subsystems are interacting with each other, i.e. inputs and outputs are connected according the connectivity equation (4), it is necessary to put it into an input-output form. For this purpose, each subsystem $i = 1, \dots, n$ classifies its inputs u_i, state variables x_i, algebraic variables a_i and outputs y_i. To conclude, this approach yields a coupled system of n semi-explicit DAEs in input-output form of (differential) index 1. For inputs u_i given by Eq. (4), find x_i, $\dot{x}_i$, a_i and y_i, such that

$$\begin{aligned} \dot{x}_i &= f_i(x_i, a_i, u_i, t) \\ 0 &= r_i(x_i, a_i, u_i, t) \\ y_i &= g_i(x_i, a_i, u_i, t) \end{aligned} \tag{5}$$

for $i = 1, \dots, n$. A careful choice of the connectivity matrix given in (4) guarantees that the coupled system obtains (differential) index 1 as well, cf. [1, 4]. E.g. one possible choice is the usage of differential states, which are not involved in any index reduction, as coupling variables.

For each subsystem (5) an arbitrary Runge-Kutta method with micro-step sizes h_i is used. The choice of the actual integration technique depends on the properties of the underlying system and can be explicit, implicit, fixed or adaptive. The whole system is integrated via a non-iterative co-simulation technique with macro-step size $H = \max(h_i)$. Further examples and explanations concerning this multirate technique can be found in [5].

3 Simulation of a Vehicle Cabin Heating via Air Conditioning

We consider a vehicle cabin heating example, cf. Fig. 1 in the system simulation software AVL CRUISE.™ M[1] Cooling and heating of the airflow and its circulation into and out of the passenger compartment is modelled using Eqs. (1)–(4). Three circuits (cooling, air flow and cabin air) are exchanging heat via heat exchangers. A fluid circuit filled with a propylene glycol water mixture is used as cooling and heating device. The fluid circuit is thermally coupled to a gas circuit (heating ventilation and air conditioning (HVAC)) via solids masses. Again, the HVAC is

[1] https://www.avl.com/de/cruise-m.

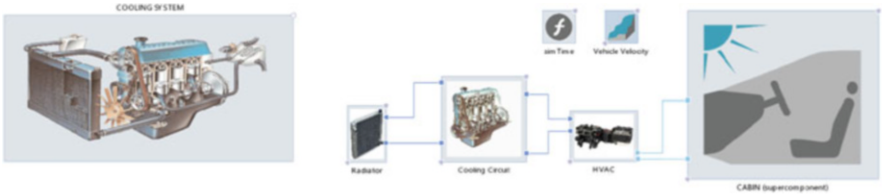


Fig. 1 Schematic representation of a vehicle cabin heating air conditioning system

Table 1 Comparison of singlerate and multirate approach corresponding CPU-time and average real time factor (RTF)

Case	CPU-time	Avg RTF	Global step	Fluid step	Gas step	Solid step	Sync step
Singlerate	177.32 s	0.5910	1ms	–	–	–	–
Multirate I	163.86 s	0.5461	–	1 ms	1 ms	1 ms	1 ms
Multirate II	143.64 s	0.4787	–	5 ms	5 ms	1 ms	1 ms
Multirate III	95.64 s	0.3188	–	5 ms	5 ms	1 ms	5 ms
Multirate IV	78.73 s	0.2624	–	10 ms	10 ms	1 ms	10 ms
Multirate V	59.03 s	0.1967	–	50 ms	10 ms	1 ms	50 ms

thermally coupled to another gas circuit (cabin air) via solid masses. Each circuit has a specific dynamic and behaviour.

This example is used to put the multirate approach presented in Sect. 2 in context with a singlerate (single solver) approach (both sequential/single CPU), cf. Table 1. The singlerate approach *Singlerate* is using one global step size of $1ms$ for all domains. On the other hand in the multirate cases the step sizes are adapted according the corresponding physical domains fluid, gas and solid. Additionally a synchronization step size is defined where the individual systems are exchanging their data. In *Multirate I-Multirate V* the step sizes of the fluid and the gas circuits are successively increased. All listed multirate cases provide a stable system simulation with sufficiently accurate physical results. Thereby, case *Multirate V* provides the most significant reduction of the calculation time from over $170s$ to under $60s$ compared to the singlerate case.

4 Conclusion

As shown, the multirate approach offers a possibility to reduce computation time considerably. In order to ensure a stable simulation, automatic index reduction of the physical networks, appropriate solver settings for each subsystem and an adequate coupling procedure, play a decisive role. For a suitable solver parametrization a significant speed up can be achieved, while conserving the accuracy criteria.

Acknowledgments Part of this work has been supported by the government of Upper Austria within the programme FTI Struktur.

References

1. Bartel, A., Günther, M.: Inter/Extrapolation-Based Multirate Schemes: A Dynamic-Iteration Perspective. Progress in Differential-Algebraic Equations II, pp. 73–90 (2020)
2. Baum, A.-K., Kolmbauer, M., Offner, G.: Topological solvability and DAE-index conditions for mass flow controlled pumps in liquid flow networks. Electr. Trans. Num. Anal., **46**, pp. 395–423 (2017)
3. Grundel, S., Jansen, L., Hornung, N., Clees, T., Tischendorf, C., Benner, P.: Model order reduction of differential algebraic equations arising from the simulation of gas transport networks. In: Progress in Differential-Algebraic Equations, pp. 183–205, Springer (2014)
4. Kolmbauer, M., Offner, G., Pöchtrager, B.: Topological Index Analysis and Its Application to Multi-Physical Systems in System Simulation Software. Progress in Industrial Mathematics: Success Stories, pp. 171–191 (2021)
5. Kolmbauer, M., Offner, G., Pfau, R., Pöchtrager, B.: Multirate DAE-Simulation and its Application in System Simulation Software for the Development of Electric Vehicles (2020) Available at https://www.ricam.oeaw.ac.at/files/reports/20/rep20-23.pdf

Mathematical Models for Electromagnetic Conditions in Submerged Arc Furnaces

Svenn Anton Halvorsen, Mads Fromreide, Manuel Sparta, and Vetle Kjær Risinggård

Abstract The competence project, “Electrical Conditions and their Process Interactions in High Temperature Metallurgical Reactors (ElMet)”, has involved close cooperation between NORCE Norwegian Research Centre, Norwegian University of Science and Technology (NTNU), the universities of Oxford and Santiago de Compostela, and the industrial companies Elkem and Eramet Norway. Various mathematical modelling has been applied to get improved insight into the inner conditions in submerged arc furnaces. Equation analysis, non-dimensionalized equations, toy models and other simplified models have been utilized to acquire sound fundamental insight. The paper describes some important results based on analysis and simple models.

1 Introduction

The design and operation of smelting furnaces have been gradually improved through industrial experience, research, modern process control, new and/or improved measurements, etc. Nevertheless, due to the complexities of the processes, several process variations are not properly understood. The furnace center is hot, above 2000 °C for some processes, and reliable measurements of the inner conditions are extremely difficult.

Smelting processes are energy intensive. The power is normally supplied by high, 3-phase, electric currents, often more than 100 000 amps. The current paths depend on electrical resistivity of the raw materials, intermediate reaction products, and the metal that is produced. As a result, there is a strong interaction between electrical current paths, temperature distribution, and chemical reactions. In a knowledge-building project, “Electrical Conditions and Their Process Interactions in High-

S. A. Halvorsen (✉) · M. Fromreide · M. Sparta · V. K. Risinggård
NORCE Norwegian Research Centre, Kristiansand, Norway
e-mail: svha@norceresearch.no; mafr@norceresearch.no; masp@norceresearch.no; veri@norceresearch.no

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_51

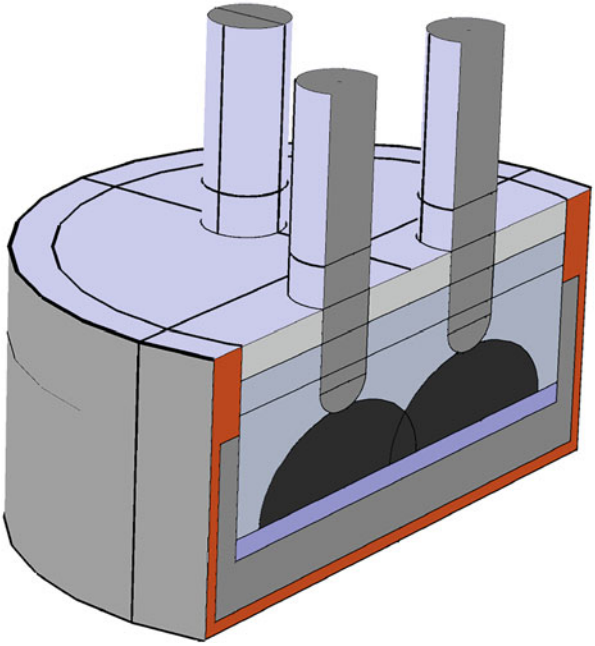


Fig. 1 Typical geometry for 3D simulations. The furnace is encapsulated by a steel shell and consists of regions with very different electrical conductivities: Raw and partly reacted materials, vertical electrodes in contact with carbon enriched regions, metal layer, and lining (carbon and ceramic brics)

Temperature Metallurgical Reactors (ElMet)", mathematical modelling has been applied to investigate this challenging problem [8]. The project has involved close collaboration among NORCE Norwegian Research Centre, the companies Elkem and Eramet Norway, and the Norwegian University of Science and Technology (NTNU), the University of Oxford, and the University of Santiago de Compostela.

Figure 1 shows the geometry for a typical full, 3D simulation, cf. for instance [6, 10]. This paper will, however, discuss how equation analysis and simplified models can give good insight, and be a valuable supplement to realistic, large, 3D models.

Electromagnetics is described by Maxwell's equations. In the ElMet project, we have assumed harmonic time variation and neglected higher harmonics. Depending on the conditions, Maxwell's equations cover [7]:

- Electromagnetic waves
- High frequency alternating current (AC)
- Low/moderate frequency AC
- Direct current (DC), or AC that at any time instant looks like DC

Previous equation analysis revealed that two physical parameters are important, the electromagnetic wavelength, λ, and the skin depth, δ, see for instance [9]:

$$\lambda = \frac{c}{f} \quad , \quad \delta = 1/\sqrt{\pi f \sigma_e \mu} \tag{1}$$

where c is the speed of light, f is the frequency (normally 50 Hz), σ_e is the electrical conductivity, and μ is the magnetic permeability.

The electromagnetic wavelength will be around 6,000 km, far bigger than the size of the furnace. Hence, electromagnetic waves and the corresponding terms in the equations can safely be ignored, cf. for instance [9]. The resulting set of equations are well known as the low-frequency time harmonic Maxwell's equations [1, 4, 5]. With this approximation, the governing equation for the electric field will be:

$$\nabla^2 \mathbf{E} - \frac{2i}{\delta^2} \mathbf{E} = \mathbf{0} \tag{2}$$

Hence, there will only be one relevant material parameter, δ, for each material. The skin depth is a characteristic length scale specifying how far the electromagnetic fields will penetrate into a sufficiently thick conductor. For a single, thick conductor, AC can be approximated by DC assuming only current in a layer of thickness δ.

2 Equation Analysis and Simple Models

First, consider a simple 2D rectangular conductor bounded by $x = 0$, $x = L$, $y = 0$, and $y = H$. As boundary conditions, let the voltage be V_0 at $x = 0$, and zero at $x = L$, and let the E-field be parallel to the upper and lower boundaries ($y = 0$, and $y = H$). In non-dimensional form, Eq. (2) can be rewritten as:

$$\frac{H^2}{L^2} \frac{\partial^2}{\partial \tilde{x}^2} \tilde{\mathbf{E}} + \frac{\partial^2}{\partial \tilde{y}^2} \tilde{\mathbf{E}} - 2i \frac{H^2}{\delta^2} \tilde{\mathbf{E}} = 0, \tag{3}$$

where we have introduced the non-dimensional parameters $\tilde{x} = Lx$, $\tilde{y} = Hy$ and $\tilde{\mathbf{E}} = E_0 \mathbf{E}$. $E_0 = V_0/L$ is a typical value for the E-field.

Then assume that the geometry is long and thin, $H \ll L$. The first term in Eq. (3) will then be very small compared to the second one, and can be neglected. For industrial purposes such approximation may be meaningful if the aspect ratio H/L is only slightly less than 1/3, due to the quadratic dependence in Eq. (3). This long and thin approximation supplies an equation for the 1D y-variation of the E-field across the thin conductor, with boundary conditions $E_x(0) = E_x(H) = V_0/L$ and $E_y(0) = E_y(H) = 0$. It is a proper approximation for the middle part but is not necessarily appropriate close to each end, i.e. close to $x = 0$ or $x = L$.

For the y-component, the trivial solution $E_y(y) = 0$ satisfies Eq. (3) and the boundary conditions at top and bottom. But the equation for E_x must be considered.

$$\frac{\partial^2}{\partial \tilde{y}^2}\tilde{E}_x - 2i\frac{H^2}{\delta^2}\tilde{E}_x = 0 \tag{4}$$

The qualitative behavior depends on the non-dimensional ratio $(H/\delta)^2$. If it is very small, the second term vanishes, and E_x can vary linearly as function of y. In our case, with the same boundary condition at the top and bottom, $E_x(y) = V_0/L$, i.e. constant. This is the direct current (DC) approximation/solution.

If $(H/\delta)^2$ is very large, the second term will dominate. Dropping the first term, we get the trivial solution $E_x(y) = 0$, but this can only be a solution inside the conductor. Close to the boundary, both terms in the equation are needed to satisfy the boundary conditions. The conditions at the top and bottom do not influence each other. The current is concentrated in thin layers at the top and bottom, and the solution is well known [7]. With y increasing into the conductor:

$$E_x(y) = E_0 e^{-\frac{1+i}{\delta}y} = (V_0/L)e^{-\frac{1+i}{\delta}y} \tag{5}$$

The total current and the power will be equal to a direct current (DC) case, where the current is confined to boundary layers of thickness δ.

If $(H/\delta)^2$ is moderate, the E-field will be somewhat reduced in the middle part, with a phase shift [7].

To provide insight for AC in a furnace, we studied a comparatively simple 2D model, cf. Fig. 2. Around each electrode tip, there is a carbon enriched layer called a carbon bed. The model includes electrodes, a carbon bed, very conductive metal below, and low conductive (partly reacted) raw materials above. Outside these materials, the model only assumes insulation ("air") [6].

The electric field is described by Eq. (2), but with a different value of δ for each material. The equation can be made non-dimensional similarly to Eq. (3). When each material is considered one by one, it follows that the electrodes and the metal behave as long and thin conductors with current confined to boundary layers characterized by δ. The raw material region is characterized by $(H/\delta)^2 \ll 1$, i.e. "DC case", while $(H/\delta)^2$ will be moderate for the coke bed.

Generalizing the insight from the rectangular conductor, the qualitative behavior of the 2D model will depend on non-dimensional parameters like $\left(H_i/L_j\right)^2$ and $(H_i/\delta_k)^2$; i.e. squared aspect ratios and squared ratios of a geometric length/height and a skin depth. The indices indicate that there are many choices, cf. Fig. 2.

Several simulations have been performed varying the relevant non-dimensional parameters. A typical result is shown in Fig. 3, where the effect of the electrode tip position is studied. The total current is the same in the two cases. The power distribution close to the electrode is largely influenced by the electrode position, while there is a large region between the electrodes where the power (and hence also the current) distributions are equal. The figure clearly shows a unidimensional, vertical variation here. The simulations revealed a strong proximity effect between the current in the metal and the parallel currents in more resistive layers above [5, 6].

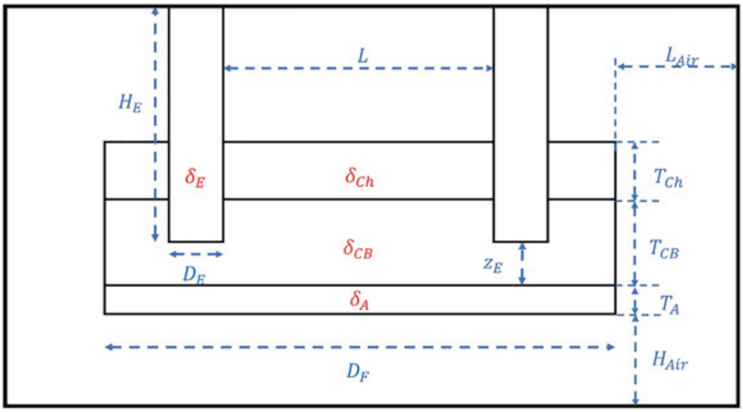


Fig. 2 Geometry of simple 2D furnace model with dimensions and material parameters indicated

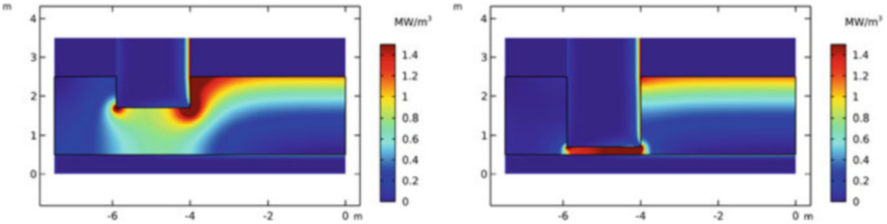


Fig. 3 2D case study: Power distributions for high and low positions of the electrodes. The figure shows the left side of the symmetric solutions

DC simulations, with current confined to boundary layers in good conductors, have been compared to AC. Significant differences were found for power in the coke bed due to *horizontal* currents, but only minor differences for *vertical* [6].

The behavior has been studied by a 1D model for AC in parallel layers [5]. For DC, the current is proportional to the conductivity in each layer (assuming a given voltage drop to drive the currents). For AC, the electric currents will be “pushed away” from a very conductive layer, into more resistive adjacent layers [4].

DC simulations are much simpler than AC and have traditionally been applied for large electric furnaces, cf. for instance [2, 3, 12, 13]. Two DC computations can also be combined into a 3D AC solution, neglecting induction effects [9]. In regions of high conductivity, the current should be confined to skin layers. Our 2D simulations show that this is not sufficient. The level and distribution of *parallel* currents in adjacent layers, will be wrong. But if the associated power is sufficiently small, DC computations might still provide a reasonable power distribution. Our 1D and 2D models can be applied to check if this is the case.

High electric currents will be induced in the (magnetic) steel shell surrounding the furnace [10, 11]. The 1D model shows that parallel currents then will be enhanced in a possible conductive (carbon) lining [4]. 3D AC simulations are recommended to check whether significant power is generated in regions where it is not wanted.

The 2D model and the 1D model for parallel layers, are proper toy models. They are too simplified for realistic (quantitative) predictions but contain essential properties of the electrical conditions. They are therefore suited to study fundamental problems. Basic understanding is far more easily detected/revealed than if more complex models are applied. The simple models can also show where realistic 3D (or 2D) simulations will be required and indicate the parameter range of interest.

3 Conclusions

During the ElMet project, we have experienced how mathematical equation analysis, non-dimensional equations, toy models and other simplified models can be applied to acquire valuable insight. Such models are very valuable, but can be far too simplified for direct, realistic, predictions. When quantitative information is needed, the simple models need to be supplemented by realistic 2D and/or 3D simulations.

Acknowledgments This work has been part of the project Electrical Conditions and their Process Interactions in High Temperature Metallurgical Reactors (ElMet), project no. 247791, with financial support from The Research Council of Norway and the companies Elkem and Eramet Norway. It is also partly financed by the project Electrical Conditions in Submerged Arc Furnaces - Identification and Improvement (SAFECI), project no. 326802, with financial support from The Research Council of Norway and four Norwegian metallurgical companies.

References

1. Bermúdez, A., Gómes, D., Salgado, P., Mathematical Models and Numerical Simulations in Electromagnetism. Springer International Publishing, 2014.
2. Darmana, D., Olsen, J.E, Tang, K., Ringdalen E. Modelling Concept for Submerged Arc Furnaces. Ninth International Conference on CFD in the Minerals and Process Industries, Melbourne, Australia, 10–12 December 2012, N.62
3. Dhainaut, M., Simulations of the Electric Field in a Submerged Arc Furnace, Infacon X, Tenth International Ferroalloy Congress, Cape Town, South Africa, 1–4 February 2004, pp. 605–613.
4. Fromreide, M., Mathematical Modelling of the Electric Behaviour of Submerged Arc Furnaces, PhD Thesis, University of Santiago de Compostela, 2022.
5. Fromreide, M., Gómez, D., Halvorsen, S.A., Herland, E.V., Salgado, P., Reduced 2D/1D Mathematical Models for analysing Inductive Effects in Submerged Arc Furnaces, Appl. Math. Model., vol. 98 (2021), 59–70.
6. Fromreide, M., Halvorsen, S.A., Sparta, M., Risinggård, V.K., Gómez, D., Salgado, P., Herland, E.V., Effects of Alternating Currents in the Hearth of Submerged Arc Furnaces, Infacon XVI, Sixteenth International Ferroalloy Congress, Trondheim, Norway, 26–29 September, 2021.
7. Griffiths, D.J., Introduction to Electrodynamcics, 3rd ed., Pearson, 2008
8. Halvorsen, S.A., Sparta, M., Risinggård, V.K., Fromreide, M., Electrical Conditions in 3-phase Submerged Arc Furnaces: Learning from the ElMet project, Infacon XVI, Sixteenth International Ferroalloy Congress, Trondheim, Norway, 26–29 September, 2021.

9. Halvorsen, S.A., Olsen, H.A.H., Fromreide, M. 2016, An Efficient Simulation Method for Current and Power Distribution in 3-Phase Electrical Smelting Furnaces. IFAC-PapersOnLine 49.20, pp.167–172.
10. Herland, E.V. Sparta M., Halvorsen S.A., 3D-models of proximity effects in large FeSi and FeMn furnaces. J. S. Afr. Inst. Min. Metall. vol. 118 (2018), 607–618.
11. Herland, E.V. Sparta, M., Halvorsen, S.A., Skin and Proximity Effects in Electrodes and Furnace Shells. Metall Mater Trans B, vol. 50 (2019), 2884–2897.
12. Karalis, K.T., Karkalos, N., Cheimarios, N., Antipas, G.S.E., Xenidis, A., Boudouvis, A.G., A CFD analysis of slag properties, electrode shape and immersion depth effects on electric submerged arc furnace heating in ferronickel processing, Appl. Math. Model., vol. 40 (2016), 9052–9066.
13. Tesfahunegn, Y.A., Magnusson, T., Tangstad, M., Saevarsdottir, G., Effect of electrode shape on the current distribution in submerged arc furnaces for silicon production - A modelling approach, J. S. Afr. Inst. Min. Metall. vol.118, (2018), 595–600.

Contaminant Removal by Adsorption

Marc Calvo-Schwarzwalder, Abel Valverde, Francesc Font, Maria Aguareles, and Timothy G. Myers

Abstract We develop a mathematical model for filtration in a cylindrical column packed with a porous material. The base model involves coupling an advection-diffusion equation to a sink term which represents the sorption and is appropriate when trace quantities are removed from the fluid. This is then extended to include the variation of velocity and pressure, which is appropriate for the removal of significant quantities, and leads to a system of five coupled equations. For the case of CO_2 removal we are able to reduce the complexity of the equations and to derive an analytical expression for the breakthrough curve. This expression is then verified against experimental data for the adsorption of CO_2 from gas and antibiotics from water. Finally, we show how the work may be modified to deal with certain extraction processes, where a clean fluid is used to remove material from the porous matrix, such as lanolin from wool.

1 Introduction

Sorption is one of the standard methods for contaminant removal. It may be applied to both liquids and gases, is regarded as efficient and relatively easy to incorporate

M. Calvo-Schwarzwalder
Zayed University, Abu Dhabi, UAE
e-mail: marc.schwarzwalder@zu.ac.ae

A. Valverde · T. G. Myers (✉)
Centre de Recerca Matemàtica, Barcelona, Spain
e-mail: avalverde@crm.cat; tmyers@crm.cat

F. Font
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: francesc.font@upc.edu

M. Aguareles
Universitat de Girona, Girona, Spain
e-mail: maria.aguareles@udg.edu

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_52

into an industrial production chain. In practical situations, column sorption is one of the most popular methods and has proved effective in a wide range of processes, such as the removal of emerging contaminants, volatile organic compounds, CO_2, dyes and salts [1].

As discussed in [2], currently accepted mathematical models present a number of errors and inconsistencies. These problems have propagated through the literature, for example the wrong choice of adsorbate density, an incorrect averaging of the equations, or inconsistent retention or neglect of terms.

In this work we derive a mathematical model consisting mainly of an advection-diffusion equation and a kinetic equation describing the mass transfer. The model is then applied to two different removal processes in Sects. 3.1 and 3.2 and the necessary modifications to apply it also on extraction processes is discussed in Sect. 3.3.

2 Mathematical Model

A schematic of the experimental set-up can be observed in Fig. 1. Due to the complex cross-sectional configuration of the column (randomly packed with sorbent material)the model is radially averaged. The flow occupies a cross-sectional area $\epsilon\pi R^{*2}$, where ϵ is the porosity of the sorbent and * notation refers to dimensional quantities.

The gas, which enters the column at a constant rate u_0^*, consists of two components that are related to the density via $\rho^* = M_1^* c_1^* + M_2^* c_2^*$, with M_i^* and c_i^* being the molar mass and molar concentration of the species i. We assume that the only species being adsorbed during the process is c_1^*. At the inlet, the concentration of each species is c_{i0}^*. In practise, the main gas component is not the one being removed, therefore we assume $c_{10}^* < c_{20}^*$. Experimentally the inlet flow rate is maintained at a constant value using a flow meter. Hence the inlet pressure varies,

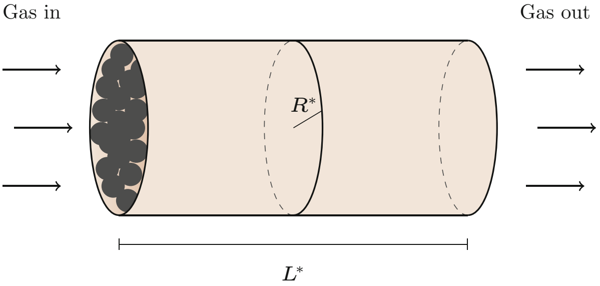


Fig. 1 A two-component gas mixture is passed through a column of length L^* and radius R^* filled with a porous adsorbing material. We assume that only one component is adsorbed in the column. The void area per unit length is described by the porosity ϵ

$p_0^*(t^*)$, whereas at the outlet the pressure is ambient p_a^*. Finally, the amount of transferred material is described by $\bar{q}^*$. Hence, the main variables of the problem are c_1^*, c_2^*, $\bar{q}^*$, u^* and p^*. Since it is often imperative to remove all contaminant for a certain period, there must exist a point within the column where $c_1 = 0$. We denote this position by $s^*(t^*)$, such that $c_1^* = \bar{q}^* = 0$ for $x^* \geq s^*(t^*)$ and consequently deal with a moving boundary problem. A standard experimental measurement is the 'breakthrough curve' this is the contaminant concentration at the outlet (which is zero until the time $s^*(t) = L^*$ where L^* is the column length).

The mass conservation equation for c_1^* includes a sink term due to the material being adsorbed. There are a number of models describing the mass transfer that defines the sink term, a very common one being the linear kinetic equation

$$\frac{\partial \bar{q}^*}{\partial t^*} = k_q^* \left(\bar{q}_s^* - \bar{q}^* \right) , \tag{1}$$

where k_q^* is the rate constant and $\bar{q}_s^*$ is a saturation value. As pointed out in [2, 3], it is common to directly integrate this equation subject to $q^*(x, 0) = 0$. This is not correct, Eq. (1) only holds in the interval $x^* \in [0, s^*(t^*)]$ and the initial condition should be $q^*(s^*(t^*), t^*) = 0$, resulting in a space-dependent expression. Finally, we close the system of equations by invoking the ideal gas law and considering a form of pressure–velocity relation derived from Navier-Stokes equations by accounting for mass loss and for viscous and inertial resistance. In the current study, we consider that the system is held at a constant temperature T^*, although we could include a heat equation if temperature variations are important.

Upon non-dimensionalisation the governing equations become

$$\delta_1 \frac{\partial c_1}{\partial t} + \frac{\partial}{\partial x}(u c_1) = \delta_2 \frac{\partial^2 c_1}{\partial x^2} - \frac{\partial \bar{q}}{\partial t} , \qquad \delta_1 \frac{\partial c_2}{\partial t} + \frac{\partial}{\partial x}(u c_2) = \delta_2 \frac{\partial^2 c_2}{\partial x^2} , \tag{2a}$$

$$\frac{\partial \bar{q}}{\partial t} = \bar{q}_s - \bar{q} , \qquad 1 + \delta_3 p = \delta_4 \left(c_2 + \delta_5 c_1 \right) , \tag{2b}$$

$$-\frac{\partial p}{\partial x} = \delta_6 (c_2 + \delta_7 c_1) u^2 + \left(1 + \delta_8 \frac{\partial \bar{q}}{\partial t} \right) u , \tag{2c}$$

with boundary and initial conditions

$$1 = \left(u c_i - \delta_2 \frac{\partial c_i}{\partial x} \right) \Big|_{x=0^+} , \qquad \frac{\partial c_i}{\partial x} \Big|_{x=L^-} = 0 , \quad i = 1, 2 , \tag{3a}$$

$$p(0, t) = p_0(t) , \quad p(L, t) = 0 , \tag{3b}$$

$$c_1(x, 0) = 0 , \quad \delta_4 c_2(x, 0) = 1 + \delta_3 p_{in} , \quad \bar{q}(x, 0) = 0 , \tag{3c}$$

where $p_{in} = p_0(0)\,(1 - x/L)$. The values of the eight non-dimensional parameters depend on each specific situation, in Sect. 3.1 we use the values from [3].

3 Applications of the mMdel

In Sect. 3.1 we apply the model on the process of carbon capture from a gas and in Sect. 3.2 we look at drug removal from water. In Sect. 3.3 we summarize how this formulation is modified to be applied to extraction and erosion processes, using the example of extracting lanolin from wool.

3.1 Removal of CO_2 from a Gas Mixture

Consider a mixture of CO_2 and N_2 flowing through a bed of activated carbon. Using the data of [4] we obtain $\delta_4 = 0.85$, $\delta_5 = 0.18$, $\delta_7 = 0.28$ whereas the remaining parameters are of the order of 10^{-2} or smaller. Neglecting small terms suggests errors of the order of 1%. Noting that $\bar{q}_s = 1 + O(\delta_3)$ [3], the problem for c_1 and $\bar{q}$ can be reduced to the coupled equations

$$\frac{\partial}{\partial x}\left(\frac{c_1}{1+\delta_{45}c_1}\right) = -\frac{\partial \bar{q}}{\partial t}\,, \qquad \frac{\partial \bar{q}}{\partial t} = 1 - q\,, \tag{4}$$

where $\delta_{45} = \delta_4\delta_5 = 0.13$. The denominator on the left results from expressing u in terms of c_1. The required boundary conditions are

$$1 = (uc_1)|_{x=0^+}\,, \qquad \left.\frac{\partial c_1}{\partial x}\right|_{x=L^-} = 0\,, \qquad \bar{q}(s(t), 0) = 0\,. \tag{5}$$

This reduced problem can be solved analytically as there exists a travelling wave solution, provided the front propagates at a constant speed. We define the variable $\eta = x - s(t)$ where $ds/dt = v$ is constant. The system can be easily be integrated and we find $v = 1$ by imposing $uc_1 \to 1$ behind the moving front (formally as $\eta \to -\infty$). Consequently $s(t) = s_0 + t$, where s_0 is determined from experimental data for breakthrough (the traveling wave solution does not hold at $t = 0$). Typically we may impose $s(t_b) = L$ where t_b is the time measured for first breakthrough or, following the discussion in [3], the time when the value of c_1 is half the inlet value, $s_0 = L - t_{1/2} + \ln(2 - \delta_{45}) = L - t_b$.

After writing the solution in dimensional form and setting $x^* = L^*$ the breakthrough curve, i.e., the concentration of CO_2 leaving the outlet, is

$$\frac{c_1^*(L^*, t^*)}{c_{10}^*} = \frac{1 - \exp\left[-k_q^*(t^* - t_b^*)\right]}{1 - (R_g^* T^* c_{10}^*/p_a^*)\exp\left[-k_q^*(t^* - t_b^*)\right]}\,. \tag{6}$$

In Fig. 2 we can see the concentration of CO_2 measured at the outlet as predicted by (6). It can be observed how our model is able to capture the trend of the experimental data with a good level of accuracy. Some qualitative differences are

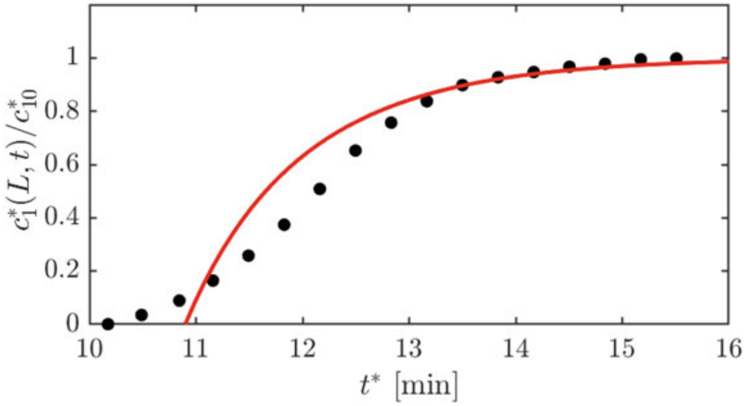


Fig. 2 Comparison between the breakthrough curve 6 and experimental data from [4], where a mixture of CO_2 and N_2 passes through a bed of activated carbon. The experimental data refers to the concentration of CO_2 measured at the outlet

observed near $t^* = t_b^*$, where the experimental data suggests a smooth increase that our model does not seem to capture.

3.2 Removal of Amoxicillin from Water

In this section we will compare the expression (6) to the available data for removal of amoxicillin in water by activated carbon [9].

First of all, we note that unlike in the case of CO_2 removal with amoxicillin we expect only trace amounts in the fluid and hence the velocity is constant everywhere. In this case the breakthrough curve reduces to a simpler form

$$\frac{c_1^*(L^*, t^*)}{c_{10}^*} = 1 - \exp\left[- k_q^*(t^* - t_b^*)\right]. \tag{7}$$

A number of alternative models can be found in the literature, such as the Thomas, Yoon-Nelson or the Bohart-Adams models [5–7]. All of these have similar expressions for the breakthrough curve,

$$\frac{c_1^*(L^*, t^*)}{c_{10}^*} = \frac{1}{1 + a^* \cdot \exp(-b^* t^*)}, \tag{8}$$

where a^* and b^* are parameters specific to each model [8]. Since their values are determined by fitting to experimental data, these models are mathematically equivalent.

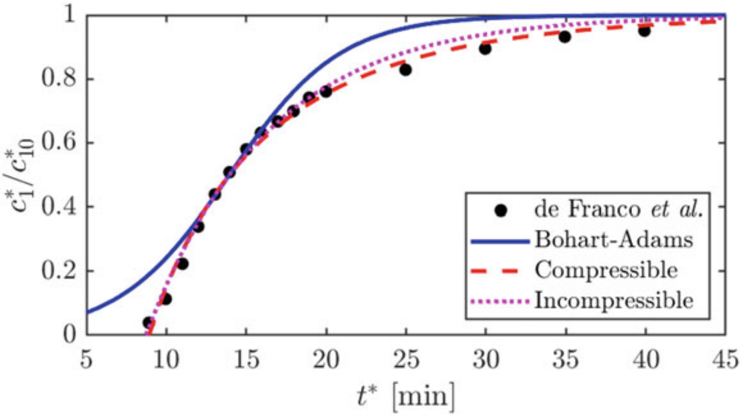


Fig. 3 Prediction of the breakthrough curve according to (6) (dashed red), (7) (dotted magenta) and (8) (solid blue). For more information about the least-square fittings we refer to [3]. Experimental data are taken from [9, Fig. 9]

The performance of the three considered formulations can be observed in Fig. 3. We can observe that both models presented here provide an excellent agreement with the experimental data, whereas (8) (which has two fitting parameters) fails to capture the trend during most of the time period considered.

3.3 Model Reformulation for Extraction Processes

Although the physics of adsorption and extraction processes are different, they can be described mathematically in a similar way. Both processes are governed by advection-diffusion equations, with the difference that the sign of the source term must be switched as material of interest is now released into the solvent rather than extracted from it. Specifically, we are interested in a situation like the one depicted in Fig. 4, where the column now contains a number of fibers formed by a solid core and an outer layer of material to be eroded. The source term is now related to the mass being released to the solvent, therefore it depends on the rate of change of the fibers' radii, which has an average value $R(x,t)$. The main advection-diffusion and mass transfer equations can be written, in non-dimensional formulation, as

$$\delta_1 \frac{\partial}{\partial t}(\epsilon c) + \frac{\partial}{\partial x}(\epsilon u c) = \delta_2 \frac{\partial}{\partial x}\left(D \frac{\partial}{\partial x}(\epsilon c)\right) - R\frac{\partial R}{\partial t}\,, \qquad \frac{\partial R}{\partial t} = -(1-c)\,. \tag{9}$$

A key difference with respect to the model for adsorption is the fact the void fraction and the diffusivity of the medium are included into the derivatives as both vary with the fiber radius R. Secondly, the source term is now nonlinear as the mass released into the solvent is proportional to R^2, hence the rate is proportional to $R(\partial R/\partial t)$.

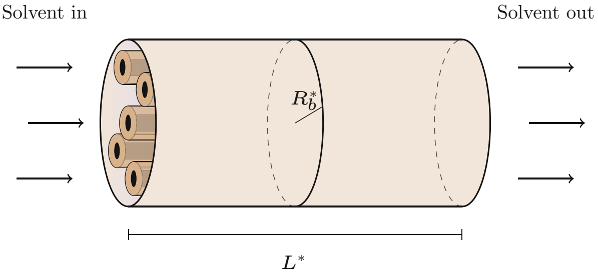


Fig. 4 Diagram showing a column of length L^* and cross-sectional radius R_b^* with variable void fraction $\epsilon(x^*, t^*)$, containing a number n of ideal cylindrical fibers with average radius $R^*(x^*, t^*)$, each consisting of a core of radius R_c^* and a outer layer of material which is eroded

To describe the decrease we have used a linear relation which states erosion occurs until the concentration in the solvent reaches a certain saturation value (here scaled to unity).

Previous investigations involve numerical solutions, see [11, 12], and none of the previous models in the literature deal consistently with the variation of the void fraction ϵ. In [13] the above equations are solved analytically by combining a perturbation method based the assumption that the total decrease of the fiber radius is small and a travelling wave. The final outcome, key to experimentalists, is the extracted fraction X, which is similar to the breakthrough curve and measures the total concentration of eroded material leaving the column outlet. In the case of lanolin, experiments show that the saturation concentration switches during the process (the newly formed outer lanolin is easily soluble, whereas the inner material tends to be much harder to remove). This requires that the constant in the R_t equation changes at a specified radius. Results from the model are compared against the experimental data for lanolin extraction of [10] in Fig. 5.

4 Conclusions

In this study we have presented a mathematical model to describe the flow of a fluid through a packed column where one of its components is being adsorbed into the medium. Approximate solutions to the model may be obtained by neglecting small terms and a travelling wave assumption. The model was verified by comparison with experimental data for the removal of CO_2 and amoxicillin. Previous formulations, such as the classical Bohart-Addams model, do not always capture the experimental data, whilst the model presented here shows an overall good agreement with it. Although there are examples where Bohart-Adams can show the better agreement. A simple adaptation of the model permitted us to also examine extraction processes.

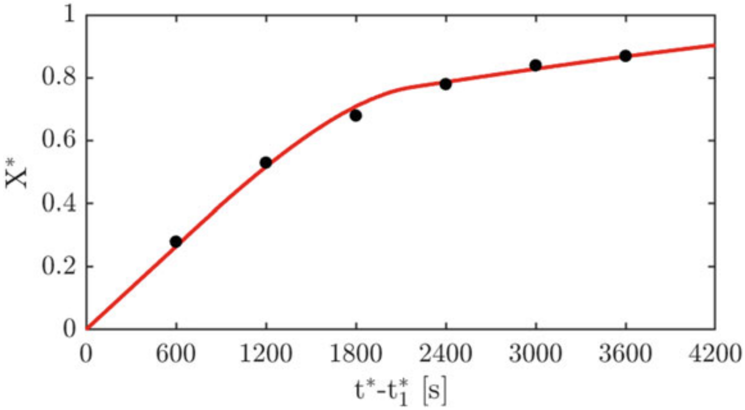


Fig. 5 Analytical prediction of extracted fraction X, taken from [13], and experimental data of Eychenne et al. [10]

Our model is not perfect, for one thing it does not capture the physics near the breakthrough time. This is the subject of our current research and, at the moment, appears to be a result of the approximation form for the contaminant sink.

Acknowledgments M. Calvo-Schwarzwalder acknowledges the funding received by the Research Office, Zayed University, UAE (Project No. R 21033). A. Valverde, T. Myers and F. Font acknowledge the support grant No. PID2020-115023RB-I00 financed by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe". A. Valverde acknowledges the support by the European Union "NextGenerationEU/PRTR" (Pruebas de Concepto Call). F. Font acknowledges financial support from the Juan de la Cierva programme (grant IJC2018-038463-I). A. Valverde, T.G. Myers and F. Font thank the CERCA Programme of the Generalitat de Catalunya. Their work was also supported by the Spanish State Research Agency, through the Severo Ochoa and Maria de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M). Francesc Font is a Serra Húnter Fellow.

References

1. Xu, Z., Cai, J.-Q., Pan, B.-C.: Mathematically modeling fixed-bed adsorption in aqueous systems. J. Zhejiang Univ.-Sci. A (Appl. Phys. and Eng.) **14**(3), 155–176 (2013).
2. Myers, T.G., Font, F., Hennessy, M.G.: Mathematical modelling of carbon capture in a packed column by adsorption. Appl. Energ. **278**, 115565 (2020).
3. Myers, T.G. Font, F.: Mass transfer from a fluid flowing through a porous media. Int. J. Heat Mass Tran. **163**, 120374 (2020).
4. Shafeeyan, M.S., Daud, W.M.A.W., Shamiri, A., Aghamohammadi, N.: Modeling of carbon dioxide adsorption onto ammonia-modified activated carbon: kinetic analysis and breakthrough behavior. Energy & Fuels **29**(10), 6565–6577 (2015).
5. Han, R., Wang, Y., Zhao, X., Wang, Y., Xie, F., Cheng, J., Tang, M.: Adsorption of methylene blue by phoenix tree leaf powder in a fixed-bed column: experiments and prediction of breakthrough curves. Desalination **245**(1–3), 284–297 (2009).

6. Yoon, Y. H., Nelson, J.H.: Application of gas adsorption kinetics–II. A theoretical model for respirator cartridge service life and its practical applications. Am. Ind. Hyg. Assoc. J. **45**(8), 517–524 (1984).
7. Bohart, G.S., Adams, E.Q.: Some aspects of the behavior of charcoal with respect to chlorine. J. Am. Chem. Soc. **42**(3), 523–544 (1920).
8. Patel, H.: Fixed-bed column adsorption study: a comprehensive review, Appl. Water Sci. **9**, 45 (2019).
9. de Franco, M.A.E., de Carvalho, C.B., Bonetto, M.M., de Pelegrini Soares, R., Feris, L.A.: Removal of amoxicillin from water by adsorption onto activated carbon in batch process and fixed bed column: kinetics, isotherms, experimental design and breakthrough curves modelling. J. Clean. Prod. **161**, 947–956 (2017).
10. Eychenne, V., Sáiz, S., Trabelsi, F., Recasens, F.: Near-critical solvent extraction of wool with modified carbon dioxide - experimental results. J. Supercrit. Fluids **21**, 23–31 (2001).
11. Valverde, A., Alvarez-Florez, J., Recasens, F.: Mathematical modelling of supercritical fluid extraction of liquid lanolin from raw wool. Solubility and mass transfer rate parameters. Chem. Eng. Res. Des. **164**, 352–360 (2020).
12. Valverde, A., Recasens, F.: Extraction of solid lanoline from raw wool with nearcritical ethanol modified CO – a mass transfer model. J. Supercrit. Fluids **145**, 151–161 (2019).
13. Myers, T.G., Valverde, A., Aguareles, M., Calvo-Schwarzwalder, M., Font, F.: Modelling mass transfer from a packed bed by fluid extraction. Int. J. Heat Mass Tran. **188**, 122562 (2022).

Vector Lattice Boltzmann Equations: From Magnetohydrodynamics to Active Matter

Paul J. Dellar

Abstract We present a lattice Boltzmann algorithm for simulating magnetohydrodynamics, and extend it to simulate the Jeffery equation that describes the rotating orientations of axisymmetric particles in a dilute suspension. Both systems involve material vector fields that evolve through the curl of another vector field. Both systems thus require an underlying kinetic formulation using vector fields, in contrast to the scalar fields used in the Boltzmann equation, and in lattice Boltzmann algorithms for hydrodynamics. Simulating Jeffery's equation requires extra gradient terms that cannot be written in conservation form. These gradients are obtained locally at grid points using the non-equilibrium parts of the kinetic vector fields representing the particle orientations, and the kinetic scalar fields representing the suspending fluid. The kinetic formulation is discretised using a Strang splitting between advection to neighbouring grid points and local algebraic operations at grid points.

1 Introduction

Magnetohydrodynamics (MHD) describes the flow of electrically conducting fluids in magnetic fields by coupling the Maxwell and Navier-Stokes equations. MHD flows arise in the interiors of stars and planets, in smelting and processing liquid metals and semiconductors, and in magnetic confinement fusion reactors [5]. The MHD equations have many structural similarities with recently-developed continuum models for suspensions of rod-like particles [13, 22, 23] that are based on Jeffery's equation for a single axisymmetric particle in Stokes flow [2, 16, 17]. This work describes how a numerical method for solving the MHD equations [6] using the lattice Boltzmann approach [1, 19] can be adapted to simulate these suspensions.

P. J. Dellar (✉)
Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Oxford, UK
e-mail: dellar@maths.ox.ac.uk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_53

2 Lattice Boltzmann Hydrodynamics

The Boltzmann equation describes a rarefied gas using a single scalar field $f(\mathbf{x}, \mathbf{c}, t)$ for the number density of particles at position $\mathbf{x}$ moving with velocity $\mathbf{c}$ at time t,

$$\partial_t f + \mathbf{c} \cdot \nabla f = C[f, f]. \tag{1}$$

The left-hand side represents linear advection of f with the particle velocity $\mathbf{c}$. All nonlinearity is confined to the right-hand side. Boltzmann's binary collision operator $C[f, f]$ describes collisions between pairs of particles via a nonlocal integral operator over $\mathbf{c}$. The Navier-Stokes equations describe solutions of the Boltzmann equation that vary slowly compared to the timescale of collisions [4].

The lattice Boltzmann approach restricts $\mathbf{c}$ to a discrete set $\mathbf{c}_0, \ldots, \mathbf{c}_N$, thus replacing $f(\mathbf{x}, \mathbf{c}, t)$ with a discrete set of functions $f_i(\mathbf{x}, t)$ that evolve according to

$$\partial_t f_i + \mathbf{c}_i \cdot \nabla f_i = -\frac{1}{\tau}\left(f_i - f_i^{(0)}\right) \tag{2}$$

for $i = 0, \ldots, N$. This right-hand side models collisions through a linear relaxation on a prescribed timescale τ towards equilibria $f_i^{(0)}$ that are prescribed functions of the local fluid density ρ and velocity $\mathbf{u}$. These macroscopic quantities are given by moments of the f_i,

$$\rho = \sum_{i=0}^{N} f_i, \quad \rho\,\mathbf{u} = \sum_{i=0}^{N} \mathbf{c}_i f_i. \tag{3}$$

In 2D, the $\mathbf{c}_i$ are commonly chosen to be the nine shown in Fig. 1 with [21]

$$f_i^{(0)} = w_i \rho \left\{1 + 3\,\mathbf{u} \cdot \mathbf{c}_i + \frac{9}{2}\left((\mathbf{c}_i \cdot \mathbf{u})^2 - \frac{1}{3}|\mathbf{u}|^2\right)\right\}. \tag{4}$$

The weights are $w_0 = 4/9$, $w_{1,2,3,4} = 1/9$ and $w_{5,6,7,8} = 1/36$. From (2) we can derive the lattice Boltzmann equation for some transformed variables $\overline{f}_i$,

$$\overline{f}_i(\mathbf{x} + \mathbf{c}_i \Delta t, t + \Delta t) = \overline{f}_i(\mathbf{x}, t) - \frac{\Delta t}{\tau + \Delta t/2}\left(\overline{f}_i(\mathbf{x}, t) - f_i^{(0)}(\mathbf{x}, t)\right). \tag{5}$$

We can integrate (2) along its characteristics [14] for a time step Δt, or we can apply a Strang splitting into advective and algebraic parts that are solved separately [8].

Using a multiple-scales expansion of both the f_i and the time derivative in a small parameter $\epsilon = \tau/T$ with suitable solvability conditions, we can find solutions of (2) for which ρ and $\mathbf{u}$ evolve on a slow hydrodynamic timescale T according to

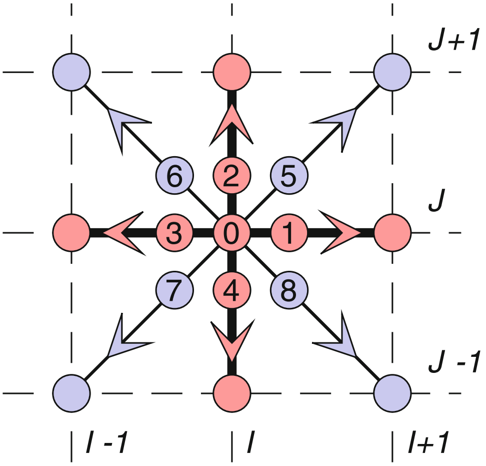


Fig. 1 The nine discrete velocities $\mathbf{c}_0, \ldots, \mathbf{c}_8$ used for the hydrodynamic scalar fields f_i and the five discrete velocities (red, thicker lines) used for the magnetic vector fields $\mathbf{g}_i$. The $\mathbf{c}_i$ are scaled so that each particle propagates from a grid point $\mathbf{x}$ to an adjacent grid point $\mathbf{x} + \mathbf{c}_i \Delta t$ over a time step. The grid points are indexed by I and J

$$\partial_t \rho + \nabla\cdot(\rho\mathbf{u}) = 0, \quad \partial_t(\rho\mathbf{u}) + \nabla\cdot(\mathit{\Pi}^{(0)} + \mathit{\Pi}^{(1)} + \cdots) = 0. \tag{6}$$

These are macroscopic mass and momentum conservation laws. The solvability conditions leave ρ and $\mathbf{u}$, the quantities conserved under collisions, unexpanded, while the momentum flux $\mathit{\Pi} = \sum_i \mathbf{c}_i\mathbf{c}_i f_i$ is expanded in ϵ as $\mathit{\Pi} = \mathit{\Pi}^{(0)} + \mathit{\Pi}^{(1)} + \cdots$.

The equilibria (4) give $\mathit{\Pi}^{(0)} = c_\mathrm{s}^2 \rho\, I + \rho\, \mathbf{u}\, \mathbf{u}$ with pressure $p = c_\mathrm{s}^2 \rho$ and constant sound speed $c_\mathrm{s} = 1/\sqrt{3}$ in "lattice units" with $\Delta x = \Delta t = 1$. We thus recover the compressible Euler equations at leading order. The multiple-scales expansion gives

$$\mathit{\Pi}^{(1)} = -\,\tau\, \rho c_\mathrm{s}^2 \left((\nabla\mathbf{u}) + (\nabla\mathbf{u})^T \right) + \tau\, \nabla\cdot(\rho\mathbf{u}\,\mathbf{u}\,\mathbf{u}), \tag{7}$$

so at next order we recover the Navier-Stokes viscous stress with dynamic viscosity $\mu = \tau\rho c_\mathrm{s}^2$, and an error term $\tau\, \nabla\cdot(\rho\mathbf{u}\,\mathbf{u}\,\mathbf{u})$. The error term is smaller than the viscous stress by the square of the Mach number $|\mathbf{u}|/c_\mathrm{s}$. It is an artifact created by using the discrete velocity set in Fig. 1 with only nine velocities.

3 Lattice Boltzmann Magnetohydrodynamics

The magnetic field $\mathbf{B}$ evolves according to Maxwell's equation $\partial_t\mathbf{B} + \nabla\times\mathbf{E} = 0$, where $\mathbf{E}$ is the electric field. We can rewrite this equation in divergence form as

$$\partial_t\mathbf{B} + \nabla\cdot\mathit{\Lambda} = 0 \tag{8}$$

using the tensor Λ with components $\Lambda_{\alpha\beta} = -\epsilon_{\alpha\beta\gamma} E_\gamma$. This now resembles the momentum equation in (6), except $\Pi = \sum_i \mathbf{c}_i \mathbf{c}_i f_i$ is symmetric by construction, while Λ is antisymmetric. It is thus impossible to represent (8) using scalar fields f_i.

Instead, inspired by work conducted at the Schlumberger-Doll laboratory to simulate magnetic resonance imaging of flow in porous media [12], we can represent the magnetic field as the sum of a set of kinetic vector fields $\mathbf{g}_i(\mathbf{x}, t)$ that evolve as

$$\partial_t \mathbf{g}_i + \mathbf{c}_i \cdot \nabla \mathbf{g}_i = -\frac{1}{\tau_\Lambda}\left(\mathbf{g}_i - \mathbf{g}_i^{(0)}\right), \tag{9}$$

where

$$\mathbf{g}_i^{(0)} = W_i \left(\mathbf{B} + \Theta^{-1} \mathbf{c}_i \cdot \Lambda^{(0)}\right), \tag{10}$$

with $\Lambda^{(0)} = \mathbf{u}\,\mathbf{B} - \mathbf{B}\,\mathbf{u}$. It is sufficient to use five discrete velocities in 2D, as shown in Fig. 1, with weights $W_0 = 1/3$, $W_{1,2,3,4} = 1/6$ and lattice constant $\Theta = 1/3$.

Summing (9) over i and applying another multiple-scales expansion gives

$$\partial_t \mathbf{B} + \nabla\cdot\left(\Lambda^{(0)} + \Lambda^{(1)} + \cdots\right) = 0, \tag{11}$$

for the moments

$$\mathbf{B} = \sum_{i=0}^{4} \mathbf{g}_i, \quad \Lambda^{(n)} = \sum_{i=0}^{4} \mathbf{c}_i\, \mathbf{g}_i^{(n)}. \tag{12}$$

We obtain ideal MHD at leading order as $\Lambda^{(0)} = \mathbf{u}\,\mathbf{B} - \mathbf{B}\,\mathbf{u}$. The first correction is

$$\Lambda^{(1)} = -\tau_\Lambda\, \Theta\, \nabla\, \mathbf{B}, \tag{13}$$

for which (11) gives the resistive MHD induction equation with resistivity $\eta = \tau_\Lambda\, \Theta$,

$$\partial_t \mathbf{B} = \nabla\times(\mathbf{u}\times\mathbf{B}) + \eta\nabla^2\mathbf{B}, \tag{14}$$

using $\nabla\cdot\mathbf{B} = 0$. The Lorentz force $(\nabla\times\mathbf{B})\times\mathbf{B}$ exerted by the magnetic field on the fluid can be rewritten as the divergence of the Maxwell stress. The Maxwell stress can be included in the equilibrium momentum flux $\Pi^{(0)}$ by redefining the $f_i^{(0)}$ [6]. The resulting algorithm has been employed for large simulations of 3D MHD turbulence [24] and of liquid metal cooling blankets for fusion reactors [20].

4 Jeffery's Equation for Axisymmetric Particles

Jeffery's equation describes a torque-free axisymmetric rigid particle immersed in a Stokes flow with a uniform velocity gradient L at infinity [2, 16, 17]. The unit vector $\mathbf{p}$ directed along the symmetry axis evolves according to

$$\dot{\mathbf{p}} = \Omega \cdot \mathbf{p} + \beta \, (E \cdot \mathbf{p} - \mathbf{p}\,\mathbf{p} \cdot E \cdot \mathbf{p}) \,, \tag{15}$$

where $\Omega = \frac{1}{2}(\,L - L^T)$ and $E = \frac{1}{2}(\,L + L^T)$ are the antisymmetric and symmetric parts of the tensor L. The last term in (15) preserves the normalisation $|\mathbf{p}| = 1$, as $\dot{\mathbf{p}} \cdot \mathbf{p} = 0$. The Bretherton shape parameter β equals $(r^2 - 1)/(r^2 + 1)$ for spheroids with aspect ratio r, so $\beta \approx 1$ for slender rods with $r \gg 1$, while $\beta = 0$ for spheres.

To describe a dilute suspension of particles we treat $\mathbf{p}(\mathbf{x}, t)$ as a vector field, and replace L with the local velocity gradient $\nabla\mathbf{u}$, assumed to vary on lengthscales much larger than the particle size. We also replace $\dot{\mathbf{p}}$ with a material time derivative,

$$\partial_t \mathbf{p} + \mathbf{u} \cdot \nabla\mathbf{p} = \mathbf{p} \cdot \nabla\mathbf{u} + (\,\beta - 1)\, E \cdot \mathbf{p} - \beta\, \mathbf{p}\,\mathbf{p} \cdot E \cdot \mathbf{p}. \tag{16}$$

To make a closer connection with the MHD induction equation, and because lattice Boltzmann algorithms simulate compressible fluids with finite sound speeds, we introduce $\mathbf{P} = \rho\, \mathbf{p}$, normalised by $|\mathbf{P}| = \rho$. The vector field $\mathbf{P}$ evolves according to

$$\partial_t \mathbf{P} = \nabla{\times}(\mathbf{u}{\times}\mathbf{P}) - \mathbf{u}\, \nabla{\cdot}\mathbf{P} + (\,\beta - 1)\, E \cdot \mathbf{P} - (\,\beta/\rho^2)\, \mathbf{P}\,\mathbf{P} \cdot E \cdot \mathbf{P}. \tag{17}$$

The first term on the right-hand side now exactly matches the MHD induction equation (14). The remaining terms arise because we have replaced $\nabla{\cdot}\mathbf{B} = 0$ by $|\mathbf{P}| = \rho$, and because particles with $\beta < 1$ do not align perfectly with the velocity gradient, in contrast to magnetic fields. However, if we represent $\mathbf{P} = \sum_i \mathbf{g}_i$ as in Sect. 3 we can obtain $\nabla{\cdot}\mathbf{P}$ from Tr Λ using (13), and E from $T = \Pi - \Pi^{(0)}$ using (7), giving

$$\partial_t \mathbf{P} = \nabla{\times}(\mathbf{u}{\times}\mathbf{P}) + \frac{1}{\tau_\Lambda \Theta}\, \mathbf{u}\ \mathrm{Tr}\ \Lambda + \frac{1-\beta}{2\tau c_s^2 \rho}\, \mathbf{P} \cdot T + \frac{\beta}{2\tau c_s^2 \rho^3}\, \mathbf{P}\,\mathbf{P} \cdot T \cdot \mathbf{P}. \tag{18}$$

5 Discretisation by Strang Splitting

To discretise the above, we separate the algebraic right-hand side of the kinetic equation for the $\mathbf{g}_i$ from the pure advection $\partial_t \mathbf{g}_i + \mathbf{c}_i \cdot \nabla\mathbf{g}_i = 0$ that gives rise to the $\nabla{\times}(\mathbf{u}{\times}\mathbf{P})$ term. The advection can be solved exactly, as in (5), and the algebraic terms by the Crank-Nicolson method. The lattice Boltzmann method relies upon an almost exact cancellation between the Crank-Nicolson truncation error and the error due to Strang splitting [3, 8].

The algebraic terms are best treated by evolving the moments of the $\mathbf{g}_i$, then reconstructing the $\mathbf{g}_i$ from the moments. For example, if we represent $\mathbf{P}$ using the five discrete velocities shown in Fig. 1, $\mathbf{P}$, Λ and $M = \sum_i \mathbf{c}_i \mathbf{c}_i \mathbf{g}_i$ form a basis of moments since $M_{\alpha\beta\gamma} \equiv 0$ for $\alpha \neq \beta$. We can reconstruct the $\mathbf{g}_i$ from

$$g_{i\beta} = \tfrac{1}{2}\left(c_{i\alpha}\Lambda_{\alpha\beta} + c_{i\gamma}c_{i\alpha}M_{\gamma\alpha\beta}\right) \text{ for } i \neq 0, \quad g_{0\beta} = P_\beta - M_{\alpha\alpha\beta}. \tag{19}$$

We can similarly complete ρ, $\rho\mathbf{u}$, Π to form a basis for the nine moments of the f_i.

The non-equilibrium momentum flux $T = \Pi - \Pi^{(0)}$ evolves under collisions as

$$\partial_t T = -(1/\tau)T. \tag{20}$$

Discretising this ODE using the Crank-Nicolson method gives

$$\frac{T(t+\Delta t) - T(t)}{\Delta t} = -\frac{1}{2\tau}\left(T(t+\Delta t) + T(t)\right), \tag{21}$$

which rearranges into

$$T' = \frac{\tau - \Delta t/2}{\tau + \Delta t/2}\, T, \tag{22}$$

on writing T' for $T(t + \Delta t)$ and T for $T(t)$. Similarly, the Crank-Nicolson method for the evolution of Tr Λ, using Tr $\Lambda^{(0)} = 0$, gives

$$\mathrm{Tr}\ \Lambda' = \frac{\tau_\Lambda - \Delta t/2}{\tau_\Lambda + \Delta t/2}\ \mathrm{Tr}\ \Lambda. \tag{23}$$

A partial Crank-Nicolson approximation for the algebraic terms in (18) is

$$\frac{\mathbf{P}' - \mathbf{P}}{\Delta t} = \frac{1}{\tau_\Lambda \Theta}\mathbf{u}\ \mathrm{Tr}\ \widetilde{\Lambda} + \frac{1-\beta}{2\tau\rho c_s^2}\mathbf{P}\cdot\widetilde{T} + \frac{\beta}{2\tau\rho^3 c_s^2}\mathbf{P}\,\mathbf{P}\cdot\widetilde{T}\cdot\mathbf{P}, \tag{24}$$

where $\widetilde{\Lambda} = \frac{1}{2}(\Lambda' + \Lambda)$ and $\widetilde{T} = \frac{1}{2}(T' + T)$. The right-hand side is evaluated using only $\mathbf{P}$, rather than a mixture of $\mathbf{P}$ and $\mathbf{P}'$. A justification for this approximation is that Tr Λ and T evolve on the fast collisional timescales τ_Λ and τ, while $\mathbf{P}$ evolves on a slow hydrodynamic timescale. Solving (24) for $\mathbf{P}'$ using (22) and (23) gives

$$\begin{aligned}\mathbf{P}' = \mathbf{P} + \frac{1}{\Theta}\,\frac{\Delta t}{\tau_\Lambda + \Delta t/2}\,\mathbf{u}\ \mathrm{Tr}\ \Lambda + \frac{1-\beta}{2c_s^2\rho}\,\frac{\Delta t}{\tau + \Delta t/2}\,\mathbf{P}\cdot T \\ + \frac{\beta}{2c_s^2\rho^3}\,\frac{\Delta t}{\tau + \Delta t/2}\,\mathbf{P}\,\mathbf{P}\cdot T\cdot\mathbf{P},\end{aligned} \tag{25}$$

where every quantity on the right-hand side is evaluated at time t.

To evolve the remaining moments Λ and M, we form the non-equilibrium part of each moment using $\mathbf{P}$, evolve the non-equilibrium part using a Crank-Nicolson time step, then reconstruct the full moment using $\mathbf{P}'$, for example

$$\Lambda' = \Lambda^{(0)\prime} + \frac{\tau_\Lambda - \Delta t/2}{\tau_\Lambda + \Delta t/2} \left(\Lambda - \Lambda^{(0)}\right), \tag{26}$$

where $\Lambda^{(0)} = \mathbf{u}\,\mathbf{P} - \mathbf{P}\,\mathbf{u}$ and $\Lambda^{(0)\prime} = \mathbf{u}\,\mathbf{P}' - \mathbf{P}'\,\mathbf{u}$. This procedure ensures that the relations (7) and (13) expressing E and $\nabla\cdot\mathbf{P}$ in terms of non-equilibrium moments hold despite $\mathbf{P}$ changing. It is equivalent to the so-called exact difference method for implementing body forces in lattice Boltzmann hydrodynamics [9, 18]. Taking the trace of (26) gives (23) as above, since $\mathrm{Tr}\ \Lambda^{(0)\prime} = \mathrm{Tr}\ \Lambda^{(0)} = 0$. From $\mathbf{P}'$, Λ' and M' we can reconstruct the post-collisional functions $\mathbf{g}_i'$ using (19), then advect them to adjacent grid points to obtain the $\mathbf{g}_i$ at the next time step,

$$\mathbf{g}_i(\mathbf{x}, t + \Delta t) = \mathbf{g}_i'(\mathbf{x} - \mathbf{c}_i \Delta t). \tag{27}$$

6 Numerical Example: Poiseuille Flow of a Suspension of Long Rods

Jeffery's equation describes how fluid flow affects the orientation of suspended particles. To obtain interesting behaviour, the particles should in turn affect the flow. Continuum models of active rod suspensions [13, 22, 23] contain a stress proportional to $\mathbf{p}\,\mathbf{p}$, equivalent to the Maxwell stress due to a magnetic field in an incompressible fluid. As a first step, we consider a suspension of passive long rods ($r \gg 1$) large enough to be unaffected by Brownian motion. The momentum flux is then [10, 11, 15]

$$\Pi = c_s^2 \rho\, I + \rho\,\mathbf{u}\,\mathbf{u} - \mu\,(2\,E + N\,\mathbf{p}\,\mathbf{p}\,\mathbf{p}\cdot E\cdot\mathbf{p})\,. \tag{28}$$

The anisotropic extra stress along $\mathbf{p}\,\mathbf{p}$ proportional to the non-Newtonian parameter $N \sim \phi r^2/\log r$ can be significant for $r \gg 1$ even at low volume fractions ϕ.

This anisotropic viscous stress is mathematically identical to a common model for the stress in a strongly magnetised plasma, known as Braginskii MHD, if we take $\mathbf{p} = \mathbf{B}/|\mathbf{B}|$ to be a unit vector parallel to the magnetic field. To obtain (28) from our kinetic formulation we adjust (22) to apply a different relaxation time $\tau_\parallel$ to the component $\mathbf{p}\cdot T\cdot\mathbf{p}$ of the stress [7]

$$T' = \frac{\tau_\perp - \Delta t/2}{\tau_\perp + \Delta t/2}\, T + \left(\frac{\tau_\parallel - \Delta t/2}{\tau_\parallel + \Delta t/2} - \frac{\tau_\perp - \Delta t/2}{\tau_\perp + \Delta t/2}\right) \mathbf{p}\,\mathbf{p}\,\mathbf{p}\cdot T\cdot\mathbf{p}. \tag{29}$$

The time $\tau_\perp$ determines the fluid viscosity perpendicular to (or in the absence of) the particles, while $\tau_\parallel = (1+N/2)\,\tau_\perp$ is enhanced by a factor of $N/2$ as in (28). We can still obtain a kinetic approximation to the strain rate from the isotropic formula

$$E = \big(T - T'\big)/(2\,\rho\, c_s^2 \Delta t), \tag{30}$$

because E is determined by the advective terms on the left-hand side of (5).

The Poiseuille flow of such a suspension, with $\mathbf{u} = u(y,t)\,\hat{\mathbf{x}}$ driven by a constant body force $f\,\hat{\mathbf{x}}$ in the usual rheological axes, is governed by the coupled system [10]

$$\partial_t \theta = -\,(1/2)\,(1-\beta\cos 2\theta)\,\partial_y u + \kappa\,\partial_{yy}\theta, \tag{31a}$$

$$\partial_t u = f + \partial_y\left(\nu(\theta)\,\partial_y u\right). \tag{31b}$$

The effective viscosity is a function of the angle θ between the rods and the x-axis,

$$\nu(\theta) = \nu_0\big(1 + N\sin^2\theta\cos^2\theta\big). \tag{32}$$

This system also includes a small orientational diffusivity $\kappa \propto \tau_\Lambda$ analogous to the resistivity in the MHD induction equation (14). For $\beta < 1$ it has solutions that oscillate in time as the rods rotate. Figure 2 shows this oscillating parabolic profile in a numerical experiment with $\beta = 0.9$, $N = 10$, and $\theta = \sin(2\pi y)$ initially.

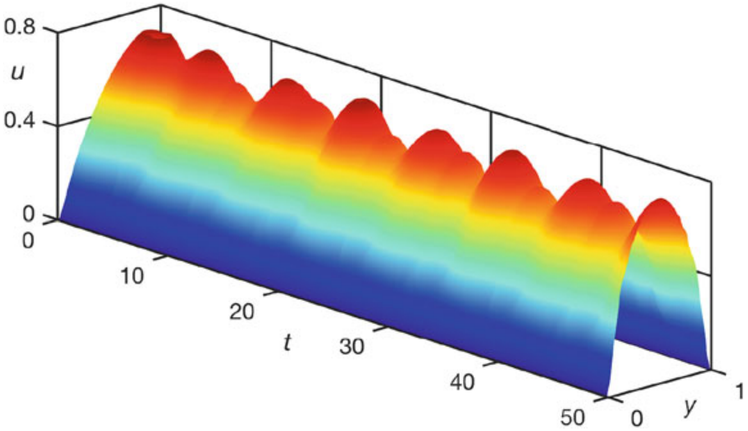


Fig. 2 Oscillatory streamwise velocity in Poiseuille flow for a suspension of elongated particles

7 Conclusion

A lattice Boltzmann approach has been presented for simulating Jeffery's equation that describes the evolution of the orientation field **p** for a suspension of axisymmetric rigid particles, exploiting its close similarity with the MHD induction equation. The primary difference is the replacement of the divergence-free constraint $\nabla\cdot\mathbf{B} = 0$ by the normalisation condition $|\mathbf{p}| = 1$. The necessary extra gradient information is available locally at grid points from the non-equilibrium parts of the hydrodynamic and orientational kinetic fields. Jeffery's equation underpins continuum models of many physical systems involving suspensions of non-spherical particles: liquid crystals, active rods, gyrotactic bacteria, and ferrofluids [22, 23].

References

1. R. Benzi, S. Succi, M. Vergassola, The lattice Boltzmann equation: theory and applications, Phys. Rep. **222**, 145 (1992)
2. F.P. Bretherton, The motion of rigid particles in a shear flow at low Reynolds number, J. Fluid Mech. **14**, 284 (1962)
3. R.A. Brownlee, A.N. Gorban, J. Levesley, Stability and stabilization of the lattice Boltzmann method, Phys. Rev. E **75**, 036711 (2007)
4. C. Cercignani, *The Boltzmann Equation and its Applications* (Springer, New York, 1988)
5. P.A. Davidson, *An Introduction to Magnetohydrodynamics*, 2nd ed. (Cambridge University Press, Cambridge, 2016)
6. P.J. Dellar, Lattice kinetic schemes for magnetohydrodynamics, J. Comput. Phys. **179**, 95 (2002)
7. P.J. Dellar, Lattice Boltzmann formulation for Braginskii magnetohydrodynamics, Comput. Fluids **46**, 201 (2011)
8. P.J. Dellar, An interpretation and derivation of the lattice Boltzmann method using Strang splitting, Comput. Math. Applic. **65**, 129 (2013)
9. P.J. Dellar, Lattice Boltzmann formulation for linear viscoelastic fluids using an abstract second stress, SIAM J. Sci. Comput. **36**, A2507 (2014)
10. J.G. Evans, The Effect of the Non-Newtonian Properties of a Suspension of Rod-like Particles on Flow Fields, in *Theoretical Rheology*, ed. by J.F. Hutton, J.R.A. Pearson, K. Walters (Applied Science Publishers, London, 1975), pp. 224–232
11. H. Giesekus, Elasto-viskose Flüssigkeiten, für die in stationären Schichtströmungen sämtliche Normalspannungskomponenten verschieden groß sind, Rheol. Acta **2**, 50 (1962)
12. R.A. Guyer, K.R. McCall, Lattice Boltzmann description of magnetization in porous media, Phys. Rev. B **62**, 3674 (2000)
13. Y. Hatwalne, S. Ramaswamy, M. Rao, R.A. Simha, Rheology of active-particle suspensions, Phys. Rev. Lett. **92**, 118101 (2004)
14. X. He, S. Chen, G.D. Doolen, A novel thermal model for the lattice Boltzmann method in incompressible limit, J. Comput. Phys. **146**, 282 (1998)
15. E.J. Hinch, L.G. Leal, The effect of Brownian motion on the rheological properties of a suspension of non-spherical particles, J. Fluid Mech. **52**, 683 (1972)
16. G.B. Jeffery, The motion of ellipsoidal particles immersed in a viscous fluid, Proc. R. Soc. Lond A **102**, 161 (1922)
17. M. Junk, R. Illner, A new derivation of Jeffery's equation, J. Math. Fluid Mech. **9**, 455 (2007)

18. A.L. Kupershtokh, Criterion of numerical instability of liquid state in LBE simulations, Comput. Math. Applic. **59**, 2236 (2010)
19. P. Lallemand, L.S. Luo, M. Krafczyk, W.A. Yong, The lattice Boltzmann method for nearly incompressible flows, J. Comput. Phys. **431**, 109713 (2021)
20. M. Pattison, K. Premnath, N. Morley, M. Abdou, Progress in lattice Boltzmann methods for magnetohydrodynamic flows relevant to fusion applications, Fusion Eng. Design **83**, 557 (2008)
21. Y.H. Qian, D. d'Humières, P. Lallemand, Lattice BGK models for Navier-Stokes equation, Europhys. Lett. **17**, 479 (1992)
22. S. Ramaswamy, The mechanics and statistics of active matter, Annu. Rev. Condens. Matter Phys. **1**, 323 (2010)
23. D. Saintillan, M.J. Shelley, Theory of Active Suspensions, in *Complex Fluids in Biological Systems: Experiment, Theory, and Computation*, ed. by S.E. Spagnolie (Springer, New York, 2015), pp. 319–355
24. G. Vahala, B. Keating, M. Soe, J. Yepez, L. Vahala, J. Carter, S. Ziegeler, MHD turbulence studies using lattice Boltzmann algorithms, Commun. Comput. Phys. **4**, 624 (2008)

A Deep Smoothness WENO Method with Applications in Option Pricing

Tatiana Kossaczká, Matthias Ehrhardt, and Michael Günther

Abstract We present the novel deep smoothness weighted essentially non-oscillatory (WENO-DS) method and its application in finance. To improve the existing WENO method, we apply a deep learning algorithm to modify the smoothness indicators of the method. This is done in a way that preserves the consistency and accuracy of the method. We present our results using a European digital option as an illustrating example. Here we avoid the undesirable oscillations, especially in the first time steps of the numerical solution.

1 Introduction

In this work, we use the newly developed weighted essentially non-oscillatory (WENO-DS) method for solving the (backward-in-time) Black-Scholes equation

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0, \quad t \in [0, T], \tag{1}$$

where S is the price of an underlying asset at time t, $r > 0$ is the riskless interest rate and σ^2 is the volatility.

The WENO method [9] is a high-order method, originally developed for solving hyperbolic conservation laws, where strong discontinuities appear in the solution. Later, it was also generalized also for solving of nonlinear degenerate parabolic equations [10]. Many modifications of the original WENO schemes have been done later and we focus in this paper on the WENO-Z method introduced in [1] and MWENO method developed in [2].

T. Kossaczká (✉) · M. Ehrhardt · M. Günther
Angewandte Mathematik und Numerische Analysis, Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: kossaczka@uni-wuppertal.de; ehrhardt@uni-wuppertal.de; guenther@uni-wuppertal.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_54

In computational finance problems, we often face the problems with discontinuous initial or terminal data. Therefore, the WENO scheme has been used, e.g. in [3, 6] for solving of these problems. In this paper, we solve the European digital option pricing problem with the following terminal and boundary conditions:

$$V(S,T) = \begin{cases} 1, & \text{if} \quad S \geq K, \\ 0, & \text{if} \quad S < K, \end{cases} \tag{2}$$
$$V(S,t) \to 0, \quad \text{for} \quad S \to 0, \quad V(S,t) \to e^{-r(T-t)}, \quad \text{for} \quad S \to \infty,$$

with K being a strike price.

Although the WENO scheme should avoid the spurious oscillations in the solution, they are still present in some cases, especially in the first time steps of the numerical solution. This motivates us to use the enhanced WENO-DS scheme [7, 8] for solving the European digital option pricing problem.

2 The WENO-DS Scheme

Here we briefly summarize the basic idea of the WENO-DS method. We consider the following diffusion-convection-reaction partial differential equation (PDE):

$$\frac{\partial u(x,t)}{\partial t} = a_0 \frac{\partial^2 u(x,t)}{\partial x^2} + a_1 \frac{\partial u(x,t)}{\partial x} + a_2 u(x,t), \quad (x,t) \in \Omega \times (0,\infty), \tag{3}$$

where a_0, a_1 and a_2 are constant coefficients. We introduce the uniform spatial grid $x_i = x_0 + i\Delta x$, $i = 0, \dots, N$. The semi-discrete formulation of (3) can be written as

$$\frac{du_i(t)}{dt} = a_0 \frac{\hat{u}_{i+\frac{1}{2}} - \hat{u}_{i-\frac{1}{2}}}{\Delta x^2} + a_1 \frac{\tilde{u}_{i+\frac{1}{2}} - \tilde{u}_{i-\frac{1}{2}}}{\Delta x} + a_2 u_i(t), \quad t > 0, \tag{4}$$

where $u_i(t)$ approximates pointwise $u(x_i,t)$ and $\hat{u}_{i+1/2} = \hat{u}(u_{i-2}, \dots, u_{i+3})$, $\tilde{u}_{i+1/2} = \tilde{u}(u_{i-2}, \dots, u_{i+2})$ are the numerical flux functions. In order to obtain these values, the WENO discretization is used.

The basic idea of the WENO scheme is to combine the numerical approximations of the flux functions on three substencils to a final numerical approximation on the main stencil. For this purpose, the nonlinear weights ω_m, $m = 0, 1, 2$, have to be calculated. For example, for the approximation of the positive part of the numerical flux of the parabolic term, one obtains

$$\hat{u}_{i+\frac{1}{2}} = \sum_{m=0}^{2} \omega_m \hat{u}^m_{i+\frac{1}{2}}, \tag{5}$$

where the explicit formulas for $\hat{u}_{i+1/2}^m$ as well as expressions of ω_m can be found in [2]. For the formulas of the numerical fluxes and the nonlinear weights for the hyperbolic term we refer to [1].

To measure the smoothness of the solution on each of three candidate substencils, the smoothness indicators $\beta_m, m = 0, 1, 2$ [4] is used. In [7] a new idea of improving these smoothness indicators was introduced. Namely they are computed as the multiplication of the original smoothness indicators β_m and the perturbations δ_m, where δ_m is an output of a particular neural network algorithm. The new smoothness indicators take the form

$$\beta_m^{DS} = \beta_m(\delta_m + C), \qquad m = 0, 1, 2, \tag{6}$$

where C is a constant that ensures the consistency and high-order accuracy of the new method, which was analytically proven in [7] and [8]. Here, also a detailed explanation of this method can be found.

3 Numerical Results

We first use the following variable transformation:

$$S = Ke^x, \quad \tau = T - t, \quad V(S, t) = Ku(x, \tau) \tag{7}$$

and substitute this into (1) and (2). Then we obtain the (forward-in-time) PDE:

$$u_\tau = \frac{\sigma^2}{2} u_{xx} + \left(r - \frac{\sigma^2}{2}\right) u_x - ru, \quad x \in \mathbb{R}, \; 0 \le \tau \le T. \tag{8}$$

This equation is of the form (3) and can be easily discretized using the WENO-DS scheme for both the hyperbolic and parabolic terms. It should be noted that for the temporal discretization, we use a third-order total variation diminishing (TVD) Runge-Kutta method, imposing intermediate boundary conditions as in [3]. Python with the Pytorch library is used for the implementation.

To obtain the enhanced WENO-DS scheme for solving the European digital option pricing problem, we train a convolutional neural network (CNN) on a large set of data. For the training, we set $K = 50$, $T = 1$, and randomly generate the parameters

$$\begin{aligned} \sigma &= 0.31 + \max(0.07a, -0.3), \\ r &= 0.11 + \max(0.07b, -0.1), \end{aligned} \tag{9}$$

where a and b are normally distributed. Here, the problems with different combinations of σ and r are covered. We use the computational domain $[x_L, x_R] = [-6, 1.5]$

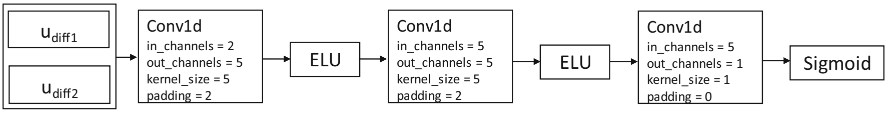


Fig. 1 The structure of the convolutional neural network

partitioned into 100 space steps and use the temporal step size $\Delta\tau = 0.8\Delta x^2/\sigma^2$. As we mentioned earlier, the spurious oscillations mainly occur in the first time steps of a numerical solution. Therefore, we proceed with a training as follows.

First, the parameters (9) are randomly generated. We initialize the weights of the CNN randomly and perform a single time step of a solution. The structure of the CNN can be seen in Fig. 1. We emphasize that we use a rather small CNN to be computationally efficient. We use the same CNN structure for training both WENO-DS for the hyperbolic term and WENO-DS for the parabolic term. We compute the values $u_{\text{diff1}}, u_{\text{diff2}}$, which represent an effective preprocessing of the solution from the current time step, since they give us information about the smoothness of the solution. They are given by

$$u_{\text{diff1},i} = \bar{u}(\bar{x}_{i+1}) - \bar{u}(\bar{x}_{i-1}), \quad u_{\text{diff2},i} = \bar{u}(\bar{x}_{i+1}) - 2\bar{u}(\bar{x}_i) + \bar{u}(\bar{x}_{i-1}), \tag{10}$$

with

$$\begin{aligned} \bar{x}_i &= (x_{i-k}, x_{i-k+1}, \dots, x_{i+k}), \\ \bar{u}(\bar{x}_i) &= \big(u(x_{i-k}), u(x_{i-k+1}), \dots, u(x_{i+k})\big), \end{aligned} \tag{11}$$

where $2k+1$ is the size of the receptive field of the whole CNN. They are then used as input values for the first hidden layer.

Then we calculate a loss with

$$\text{LOSS}(u) = \sum_{i=0}^{N-1} \big[\max(u_i - u_{i+1}, 0)\big], \tag{12}$$

where u_i is a numerical approximation of $u(x_i)$. This loss is positive, if the approximation of the solution is decreasing in x (in true solution it should be only increasing), so we test the monotonicity of the solution. After that, the gradient with respect to the weights of the CNN is calculated using the backpropagation algorithm. Then, the Adam optimizer [5] with a learning rate of 0.001 is used to update the weights. Next, we test the model on a validation set and repeat the above steps with newly generated parameters (9). After the training, we select the weights from the training step, at which the model performed best on the validation problems.

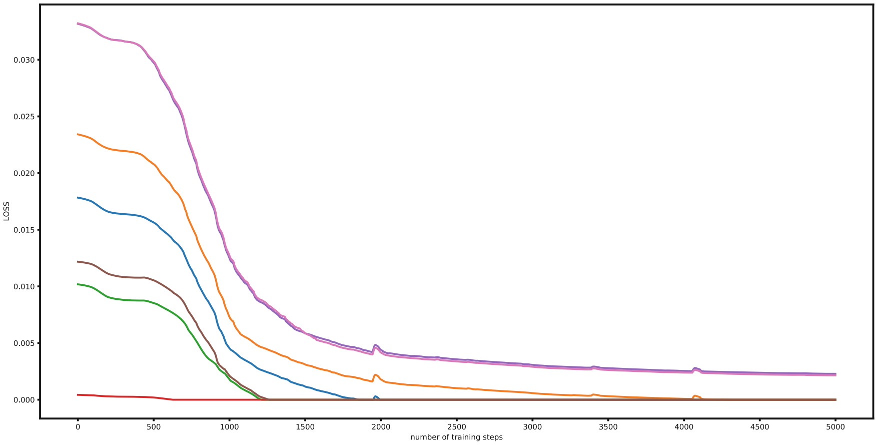


Fig. 2 Loss values for different validation problems

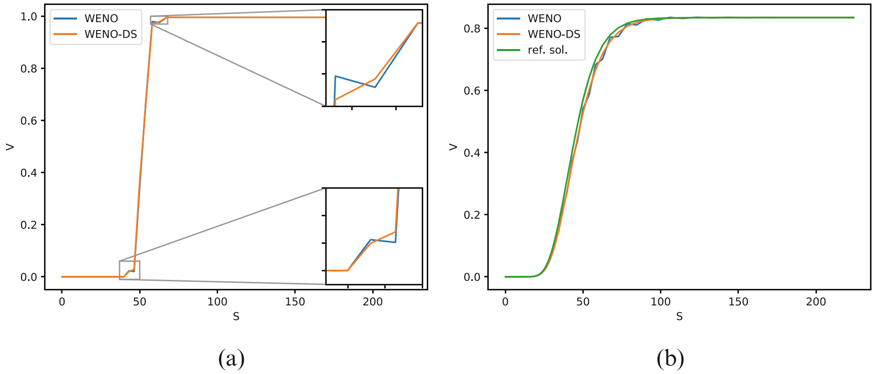


Fig. 3 Comparison of the original WENO and WENO-DS methods, $N = 100$. (**a**) Solution at the first time step, $\sigma = 0.4$ and $r = 0.15$. (**b**) Solution at the last time step, $T = 1$, $\sigma = 0.3$ and $r = 0.2$

In Fig. 2, we show the evolution of the loss value for the problems from the validation set. We see that the loss is decreasing and select the model obtained after the last training step as our final WENO-DS scheme.

We compare the solution at the first time step on Fig. 3a and see that the WENO-DS reliably eliminates the oscillations that occur when using the original WENO scheme (WENO-Z scheme [1] for the approximation of the hyperbolic term and MWENO scheme [2] for the approximation of the parabolic term).

In most cases, the original WENO scheme is able to handle these oscillations with increasing number of time steps. However, in some cases the oscillations are still present. Figure 3b shows the solution at time $T = 1$ and we see that our method produces a smooth solution unlike the original WENO method.

Table 1 Comparison of the L^∞ and L^2-error of original WENO and WENO-DS methods for the solution of the transformed Black-Scholes equation (8) with various parameters σ and r

		L^∞		L^2	
σ	r	WENO	WENO-DS	WENO	WENO-DS
0.28	0.13	0.000933	0.000908	0.000660	0.000644
0.1	0.05	0.002751	0.002655	0.001196	0.001158
0.3	0.2	0.001120	0.000858	0.000650	0.000621
0.2	0.1	0.001833	0.001687	0.000890	0.000865
0.15	0.05	0.002446	0.002352	0.001055	0.001034
0.4	0.1	0.000676	0.000661	0.000570	0.000557

We compare the L^∞ and L^2 errors in Table 1 and show that the WENO-DS method has a smaller error in all cases. Thus, we are not only able to eliminate the spurious oscillations, but also improve the quality of the numerical solution.

4 Conclusion

In this work, we applied the newly developed WENO-DS method to the European digital option pricing problem that has discontinuous terminal data. In this problem, the spurious oscillations are present in the solution when the standard WENO scheme is used. We have shown that they can be successfully eliminated using the WENO-DS method. To this end, we trained a CNN to modify the smoothness indicators of the original method. Since we can obtain smaller errors with the proposed algorithm, the quality of the numerical solution was also improved.

References

1. Borges, R., Carmona, M., Costa, B., and Don, W.S., An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws. J. Comput. Phys. 227(6) (2008), 3191–3211.
2. Hajipour, M., and Malek, A., High accurate NRK and MWENO scheme for nonlinear degenerate parabolic PDEs. Appl. Math. Model. 36.9 (2012): 4439–4451.
3. Hajipour, M., and Malek, A., High accurate modified WENO method for the solution of Black–Scholes equation. Comput. Appl. Math. 34(1) (2015), 125–140.
4. Jiang, G.-S., and Shu, C.-W., Efficient implementation of weighted ENO schemes. J. Comput. Phys. 126(1) (1996), 202–228.
5. Kingma, D.P., and Ba, J., Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014), published as a conference paper at ICLR 2015.
6. Kossaczká, T., The Weighted Essentially Non-Oscillatory Method for Problems in Finance. Master Thesis, University of Wuppertal (2019).
7. Kossaczká, T., Ehrhardt, M., and Günther, M., Enhanced fifth order WENO shock-capturing schemes with deep learning. Results Appl. Math. 12 (2021), 100217.

8. Kossaczká, T., Ehrhardt, M., and Günther, M., A neural network enhanced WENO method for nonlinear degenerate parabolic equations, Physics of Fluids 34(2) (2022), 026604.
9. Liu, X.-D., Osher, S., and Chan, T., Weighted essentially non-oscillatory schemes. J. Comput. Phys. 115(1) (1994), 200–212.
10. Liu, Y., Shu, C.-W., and Zhang, M., High order finite difference WENO schemes for nonlinear degenerate parabolic equations. SIAM J. Sci. Comput. 33(2) (2011), 939–965.

Proactive Dengue Management System Synergize by an Exponential Smoothing Model

W. A. U. K. Wetthasinghe, A. M. C. H. Attanayake, U. P. Liyanage, and S. S. N. Perera

Abstract In a critical area like health sector centralized computer system helps to improve the efficiency of the health system. In particular, controlling an epidemic is usually difficult in developing countries. In this study we introduce a multi-platform, centralized pro-active management system to manage dengue controlling activities in Sri Lanka. The system make common platform (ProDMS) for all sectors who contribute their services for mitigating dengue. We mainly focused to the special feature of the system which enhance the centralized property. Cross platform environment was developed under this feature as a bridge to connect researches and general public. ProDMS is a internet base web application and researches can plug their dengue forecasting models to the system and publish their outputs as graphs through the web system. The ProDMS web application, which consisting of plug and play system architecture concepts, fully support for any statistical or mathematical model to publish its results online. In this work we use one of the univariate time series modelling approaches; namely exponential smoothing to plug with the system. This research helps to enhance efficiency of Dengue controlling process and support to generalize centralization.

1 Introduction

Information centralization is a prevalent technique in information technology and also plays a major role in technological development in the present world in many

W. A. U. K. Wetthasinghe (✉) · S. S. N. Perera
Research & Development Centre for Mathematical Modelling, Department of Mathematics, University of Colombo, Colombo, Sri Lanka
e-mail: udana@maths.cmb.ac.lk; ssnp@maths.cmb.ac.lk

A. M. C. H. Attanayake · U. P. Liyanage
Department of Statistics and Computer Science, Faculty of Science, University of Kelaniya, Kelaniya, Sri Lanka
e-mail: succ@kln.ac.lk; liyanage@kln.ac.lk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_55

fields. Having information in a centralized location would be advantageous to users to make illustrations and their associated conclusions rapidly. Transparency, productivity, Efficiency and centralized information are the advantages along with many others that motivate the implement of such centralized system in any real situation. In particular, the information systems in the health sector require centralized operational possibilities and the quick access of information to make decisions as quickly as possible. Addressing these advantages, we developed a centralized system for dengue management.

Dengue is a vector-borne viral infection that transfers by the bite of an infected mosquito. In last five decades, dengue has reached to the level of hyper epidemic decease state in Sri Lanka [1]. Recent statistics illustrates 105049 suspected dengue cases for the year 2019 and 27886 cases in year 2020, making dengue is still the major threat in Sri Lanka [2].

In the absence of a fully effective vaccine or treatment [3], controlling vector density is the only mitigating technique that leads to the prevention. Most of dengue epidemic strategies are based on reactive system rather than pro-active system. National Dengue Control unit, Epidemiology Unit, government and privet hospital administrators, provincial/regional health officers and other corresponding parties apply appropriate integrated vector management strategies to control the spread of the disease transmission [4]. Hence, there is a need of general infrastructure or mechanism that allows efficient collaboration among all contributing parties. Therefore, in this study, such a system called “Proactive Dengue Management System (ProDMS)” is proposed to have efficient dengue management and prevention.

The proposed System (ProDMS) provides an integrated environment to all the stakeholders: policy-makers, researchers and general public. ProDMS platform is a cross-platform GIS based distributed system for monitoring, forecasting and controlling the dengue epidemics. Since various types of factors influencing the spread of the disease, mathematical and statistical models play a major role in identifying and forecasting the dynamics of the dengue disease. Further, the system will function as a dengue early warning system, which allows authorities to predict and control dengue epidemics before it rises to peak levels; thereby, efficiently controls the dengue epidemics.

The other commonly existing systems provide facilities to integrate the dengue models with management systems that are developed on same platforms, e.g. DengueME [5]. However, up to the authors’ knowledge, they have no ability to make connections between applications that are developed on different platforms. In such scenarios, researchers are unable to develop and simulate the results of their models in the comfort and specific platforms. In the ProDMS, the plug and play architecture is used to overcome such weaknesses. Further, using methods that allows to communicate and synergies features across platforms, the authors introduce a novel concept to expand the use of computational mathematics and statistic in the process of efficient decision making towards the dengue dynamics.

2 Pro-Active Dengue Management System (ProDMS)

The main objective of the ProDMS is develop the system to provide an integrated environment to the health policy makers, researchers and general public to synergize the prevention of Dengue epidemics through sharing information, knowledge enhancements, collaborations and community engagements [6].

The first prototype of the ProDMS platform and its plug and play architecture have been tested by incorporating a exponential smoothing (statistical) model that forecasts dengue dynamics. However, the current prototype of the ProDMS enables to visualize the result of its plugged the statistical model. Further, the system now allows the public and policy holders to perform different action towards dengue management related activities. The research component of this work is introducing a system architectural design that communicates dengue forecasting models with the web based application.

3 System Architecture and Implementation

Statistical model integration process is comprised of three components namely, central database, web based management system and dengue modeling tools. In the standard development mechanisms, statistical models and web based platforms are developed using different languages as well as environment. Therefore, the results may not be presented in a single environment. Herein, we introduce a steady system architecture and use a plug and play independent agent to make connections among the components. All links of the system build with data ropes. In this study, we use JavaScript Object Notation (JSON) data carrier as the data rope.

The properties of JSON: human readable text and language independence, were utilized in the stability enhancing of the cross-platform data integration. Main web based system share the dengue related data that exists in the data management system through REST API. The REST API is designed to grab advantage of existing protocols. REST can be used over any protocol, but usually takes advantage of HTTP when it comes to web APIs, therefor, in ProDMS as well. Fielding and Taylor introduced the REST API design [7]. Since the data is not integrated with methods and resources, the REST API enhances the plug and play architecture [8].

Further, the JSON data structures is defined according to centralized system database structure. This allows the third party users to request the data through the system in a secure manner through the REST API. Since JSON is language independent data format, it is compatible with most of the scientific simulation platforms. Figure 1 describe the idea of the connectivity among system components.

In this work R software is used to implement the statistical model. R is one of widely used statistical programming platform that supports the JSON data carrier, [9]. Within the plug and play architecture, the ProDMS web based application is featuring the connectivity with the statistical models developed in R, while JSON

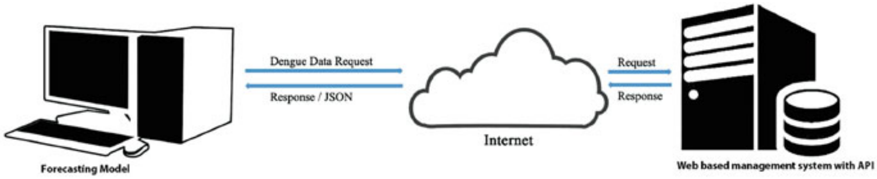


Fig. 1 Structure of statistical model integration process

is bridging the data communication. However, the data resolution at the ProDMS's stored data, e.g., weekly or monthly, have to be tallied with the data requirements of the models developed in R. The simulation results generated by statistical models are communicated through the integrated REST API using JSON data carrier. This enables an independent system connectivity among dengue model and database, and consequently, a secure system.

Currently system offer reported dengue cases, rainfall, maximum and minimum temperature and humidity under the monthly resolution. Among them, the statistical model use only reported dengue cases as input the parameter of the model. Therefore, the model filter the JSON data and extract relevant data fields for the model.

The important phase of the process exhibits the simulation results via web application. In this task, forecasting models communicates their outputs using JSON format according to given structure. These JSON structures are provided by the ProDMS API for anyone to use. Further, model passes the configuration details via output JSON. Followings are the information that should include in the JSON to use ProDMS platform.

- **Model behavior:** This property use to identify the model continuity. Model should mention that whether the model is running on "continuous time" or "discrete time". In the case of continuous time, ProDMS automatically scale and interpolate appropriate values when they are being displayed. For the discrete time scales, no adjustment will be made. This ensures the agility of the systems for any type of output given by the external models.
- **District:** In representing district related dengue dynamics, the external Model should mention the district of the output data so that it would be correctly displayed.
- **Time period:** Model should mention the time period which is used for the model simulation.

The ProDMS support for visualizing dengue data for its users, such as island-wide epidemics, summery results, forecasting information, and etc..., supported by Google-API. Further, the model results can be interpreted using infrastructures of the ProDMS system. Data are entered from dengue forecasting models can be exhibit in meaningful ways by using the system features. Researchers can use that facility to exhibit their findings to public via more general and popular way using

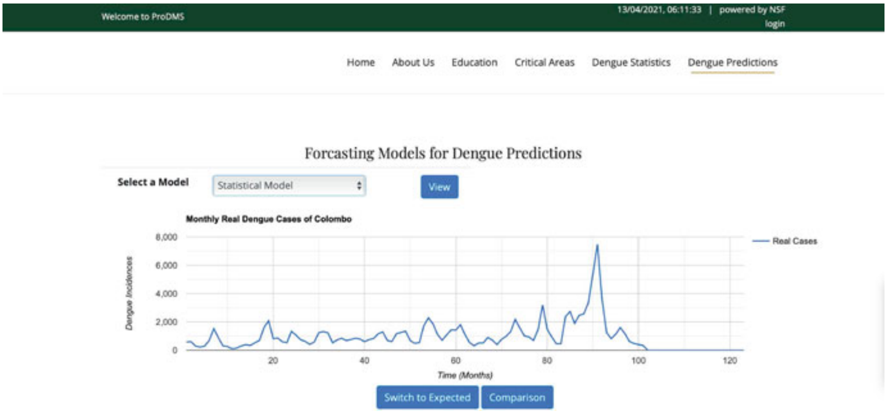
ProDMS web application. This enables users and policy makers to get a wide view of the ongoing epidemics and lead to better management.

4 Results and Discussion

4.1 Testing

One may find different mathematical and statistical models in literature such as compartmental models and regression models addressing dengue transmission [10–14]. For this study we perform some testing with the exponential smoothing model developed by our research group in previous studies [10] with use of Colombo Municipal Council (CMC) data. Figure 2 exhibits the actual dengue data which are existing in the system database.

Users can generate graphs in real time based on data of exponential smoothing model and compare with real data. Since we formulated the exponential smoothing model based on CMC data, in this study we offered only CMC data. In exponential smoothing, recent observations are weighted more heavily than older observations. Simple, double or Holt winters exponential smoothing model will be fitted to the data according to the seasonal and trend structures present in the data. 'ets' function of the 'forecast' package which is available in R language is used to fit the appropriate exponential smoothing model. Figure 3 illustrates the comparison of R simulation results and actual data.



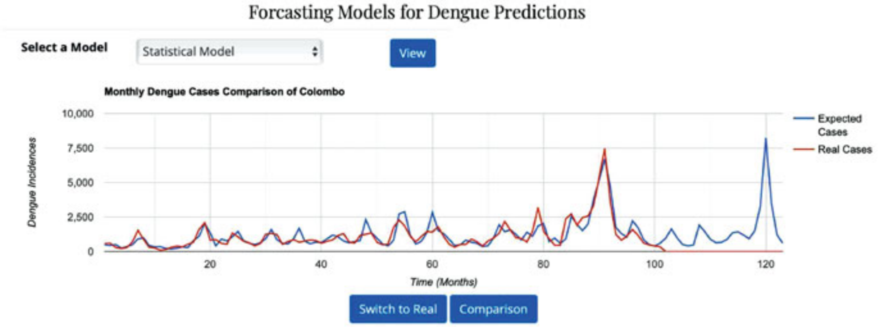
Welcome to ProDMS

13/04/2021, 06:11:33 | powered by NSF

Forcasting Models for Dengue Predictions

Select a Model: Statistical Model

Fig. 2 Reported dengue data from, CMC area



Forcasting Models for Dengue Predictions

Select a Model: Statistical Model

Fig. 3 Comparison of model results and actual data

4.2 *Future Direction*

The architecture allows future extensions to synergize more mathematical, statistical or hybrid models. Then the users can choose the appropriate model for their purpose and simulate results. Moreover, this system can be directly use to manage other epidemic situations such as dengue and COVID-19.

5 Conclusion

In this study, we introduced a plug and play system architectural design to develop infrastructure to synergize dengue forecasting models with the web based application. By this architecture, the connectivity among ProDMS, Databases and External models is archived with minimum dependencies. Consequently, a better communication among researchers and the public society, and hence, better dengue management. The systems architecture itself allows researches to use the available resources to test their modules, motivating the collaborative research activities that involve different components. Further, this study has provided a platform to researches to present their finding directly to the general public inspiring the commercialized research outcomes.

Acknowledgments This work was partially supported by National Science Foundation, Sri Lanka. Grant number RPHS/2016/D/05.

References

1. Sirisena, P.D.N.N. and Noordeen, F., Evolution of dengue in Sri Lanka-changes in the virus, vector, and climate, Int.J. Infectious Diseases, 19, 6–12, 2014.
2. National Dengue Control Unit, Ministry of Health, Nutrition and Indigenous, Intensive inter sectoral programme for the prevention and control of dengue, 2018.

3. Idris, F., Ting, D.H.R. and Alonso, S., An update on dengue vaccine development, challenges, and future perspectives. Expert Opinion on Drug Discovery, 16(1), 47–58, 2021.
4. Thalagala, N., Tissera, H., Palihawadana, P., Amarasinghe, A., Ambagahawita, A., Wilder-Smith, A., Shepard, D.S. and Tozan, Y., Costs of dengue control activities and hospitalizations in the public health sector during an epidemic year in urban Sri Lanka. PLOS Neglected Tropical Diseases, 10(2), e0004466, 2016.
5. De Lima, T.F.M., Lana, R.M., De Senna Carneiro, T.G., Codeço, C.T., Machado, G.S., Ferreira, L.S., De Castro Medeiros, L.C. and Davis Junior, C.A., Dengueme: A tool for the modeling and simulation of dengue spatiotemporal dynamics. Int. J. Environm. Res. Public Health, 13(9), 920, 2016.
6. Wetthasinghe, W.A.U.K., Liyanage, U.P. and Perera, S.S.N. Multiplatform dengue management android mobile application. AIP Conference Proceedings, 2184(1), 060025, 2019.
7. Fielding, R.T. and Taylor, R.N., Architectural styles and the design of network-based software architectures, 7. University of California, Irvine, 2000.
8. MuleSoft, What is a RESTful API?. Url: https://www.redhat.com/en/topics/api/what-is-a-rest-api (accessed: 24.01.2020).
9. Ooms, J., The jsonlite package: A practical and consistent mapping between json data and R objects. arXiv preprint arXiv: 1403.2805, 2014.
10. Attanayake, A.M.C.H., Perera, S.S.N. and Liyanage, U.P., Exponential smoothing on forecasting dengue cases in Colombo, Sri Lanka. Science, Eastern University, Sri Lanka, 11(1), 11–22, 2020.
11. Bhuju, G.,, Phaijoo, G.R. and Gurung, D.B., Fuzzy approach analyzing SEIR–SEI dengue dynamics. BioMed Research Int., 2020, 2020.
12. Erandi, K.K.W.H., Perera, S.S.N. and Mahasinghe, A.C., Analysis and forecast of dengue incidence in urban Colombo, Sri Lanka, Theor. Biol. Med. Model., 18(1), 1–19, 2021.
13. Ganegoda, N.C., Götz, T. and Wijaya, K.P., An age-dependent model for dengue transmission: Analysis and comparison to field data. Appl. Math. Comput., 388, 125538, 2021.
14. Johansson, M.A., Dominici, F. and Glass, G.E., Local and global effects of climate on dengue transmission in Puerto Rico. PLOS Neglected Tropical Diseases, 3(2), e382, 2009.

Multipatch ZIKV Model and Simulations

Arsha Sherly and Wolfgang Bock

Abstract In this article we compare two multi-patch models for the spread of Zika virus based on an SIRUV model. When the commuting between patches is ceased we expect that all the patches follow the dynamics of the single patch model. We show in an example that the effective population size should be used rather than the population size of the respective patch.

1 Introduction

Zika Virus belongs to the family *Flaviviridae*, genus *Flavivirus*. ZIKV disease is primarily vector-borne, which is transmitted by *Aedes* mosquitoes [1]. This disease is also found to be sexually transmissible [2]. Eventhough most patients show mild symptoms, recent studies show that this virus attack results in neurological disorders like Guillain-Barré syndrome (GBS) [3]. Another important characteristic of this virus is its pathogenicity to fetuses causing Microcephaly in newborn babies [4].

The history of ZIKV disease known so far starts with the isolation of Zika virus from a rhesus monkey in Uganda around April 1947. There onwards it has spread across the world with the largest outbreak recorded in 2015-16 across South America [1, 5]. With no vaccines or medications found so far, the disease spread can only be controlled by non-pharmaceutical interventions. Also increased international travel, evolution and mutation of viruses and their transmitting agents like mosquitoes, suitable environmental conditions etc lead to an increase in further outbreaks even in lesser probable places. The influence of human mobility plays an important role in transmitting diseases across continents. With more flight connectivity and affordable modes of transport disease transmission can also be faster. The primary objective of this study is to include spatial dependence to the mechanistic model of ZIKV spread. This is very relevant as the parameters involved

A. Sherly (✉) · W. Bock
TU Kaiserslautern, Kaiserslautern, Germany
e-mail: sherly@mathematik.uni-kl.de; bock@mathematik.uni-kl.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_56

in the model will be different for different places. So the dynamics will be exhibiting variations spatially. In this article we use an SIRUV model to describe the disease dynamics. This model divides the population into various compartments namely susceptible, infected and recovered. The interaction between various host and vector compartments, spread across different patches, is modeled using a coupling matrix and certain parameters.

We have discussed two models in Sects. 2 and 3. The results of numerical simulations are provided in Sect. 4. Comparing the two models exemplarily shows that the incorporation of the effective population size is crucial. While in a model, which just takes into account, the total population size of the patches, a decoupling does not lead to the single patch dynamics, where as a model which incorporates the effective population size shows this desired property.

2 Multi-Patch ZIKV Model

In this section we give a multi patch model for studying the ZIKV disease spread. Let the space domain be divided into small areas which we name as patches. The ZIKV model in a specific patch is also developed using different compartments. Here the host and vector population consists respectively of susceptible and infected compartments in each patch and we consider the recovered ones only in host population of each patch. We use either a subscript or a superscript (i, j or k) to distinguish these compartments and the parameters patchwise. Let us first assume that the whole population is commuting between the patches and the rate of transition from patch (i) to (j) be p_{ij}.

Remark 1 The matrix P with entries p_{ij} is the residence time budgeting matrix. Here p_{ij} represents the time spent by people in patch i on average in patch j in unit time [6]. For example on average if a person in patch i spent 8 hours in patch j, then $p_{ij} = \frac{8}{24}$, provided that unit time is one day.

We have deduced the following model from similar models in the literature used for other epidemiological studies [7].

$$\frac{dS_i}{dt} = \mu_i (1 - S_i) - S_i \Bigg(\sum_{1 \le j \le n} \beta_{vh}^j p_{ij} V_j + \Bigg(\sum_{1 \le j \le n} \beta_{hh}^i (p_{ij} + p_{ji}) I_j - \beta_{hh}^i p_{ii} I_i \Bigg) \Bigg)$$

$$\frac{dI_i}{dt} = S_i \Bigg(\sum_{1 \le j \le n} \beta_{vh}^j p_{ij} V_j + \Bigg(\sum_{1 \le j \le n} \beta_{hh}^i (p_{ij} + p_{ji}) I_j - \beta_{hh}^i p_{ii} I_i \Bigg) \Bigg) - \mu_i I_i - \gamma_i I_i$$

$$\frac{dR_i}{dt} = \gamma_i I_i - \mu_i R_i$$

$$\frac{dU_i}{dt} = \nu_i (1 - U_i) - \vartheta_i U_i \sum_{1 \le j \le n} I_j p_{ji}$$

$$\frac{dV_i}{dt} = \vartheta_i U_i \sum_{1 \le j \le n} I_j p_{ji} - \nu_i V_i .$$

3 Redefining the Model for ZIKV

Following some insights from [8] and [9] we have developed a new model to describe the ZIKV disease spread. In [9] a term called contact rate is clearly defined, which is the average number of adequate contacts per day of an infective person from patch j with any individuals in patch i. With this in consideration we redefine the parameters used as follows

α_j = number of infectious contacts that is happening per infected mosquito per unit time with the people present in patch j.
β_j= number of infectious contacts that is happening per infective individual per unit time with the people present in patch j.
γ_j = number of recoveries that is happening per unit time in patch j.
ϑ_j = number of infective contacts that is happening per infected human with mosquitoes in patch j in unit time.

Let us focus on patch j and see how many susceptibles from patch i gets infected in patch j. If N_j inhabitants are residing in patch j, they commute to other patches in unit time. So the effective population in patch j is given by $N^j_{\text{eff}} = \sum_{k=1}^n p_{kj} N_k$. By the definition of α_j, the number of people getting into adequate contacts with the mosquitoes in patch j is given by $\alpha_j \mathcal{V}_j$. The effective population of susceptibles in patch j is $\sum_{k=1}^n p_{kj} \mathcal{S}_k$ among which $p_{ij}\mathcal{S}_i$ are coming from patch i. In turn, the number of susceptibles from patch i who get infected in patch j due to mosquitoes is given by

$$\alpha_j \mathcal{V}_j \frac{p_{ij}\mathcal{S}_i}{\sum_{k=1}^n p_{kj} N_k}.$$

Now we focus on the infections between humans. The number of infections happening in patch j in unit time due to human-human interactions is given by $\beta_j I^j_{\text{eff}}$, where I^j_{eff} is the effective number of infected people who came to patch j in unit time which is given by $I^j_{\text{eff}} = \sum_{k=1}^n p_{kj} \mathcal{I}_k$. The total number of infections happening in patch j is given by $\beta_j \sum_{k=1}^n p_{kj} \mathcal{I}_k$, out of which the number of infections happened to the susceptible people of patch i is

$$\beta_j \sum_{k=1}^n p_{kj} \mathcal{I}_k \frac{p_{ij}\mathcal{S}_i}{\sum_{k=1}^n p_{kj} N_k}.$$

Now we have to introduce fractions by normalising each compartmental values. For example we define $S_i = \frac{\mathcal{S}_i}{N_i}$ or as in the vector population we have $U_i = \frac{\mathcal{U}_i}{M_i}$, where M_i is the number of vectors present in patch i.

Remark 2 For U_i the normalisation yields,

$$
\begin{aligned}
\frac{d\mathcal{U}_i}{dt} &= \nu_i(M_i - \mathcal{U}_i) - \vartheta_i \frac{\mathcal{U}_i}{M_i}\sum_{k=1}^{n} p_{ki} N_k I_k \\
\Leftrightarrow \frac{M_i dU_i}{dt} &= \nu_i(M_i - M_i U_i) - \vartheta_i U_i \sum_{k=1}^{n} p_{ki} N_k I_k \\
\Leftrightarrow \quad \frac{dU_i}{dt} &= \nu_i(1 - U_i) - \vartheta_i \frac{U_i}{M_i}\sum_{k=1}^{n} p_{ki} N_k I_k .
\end{aligned}
$$

The following system of ODEs describe disease spread in each patch i

$$
\begin{aligned}
\frac{dS_i}{dt} &= \mu_i(1 - S_i) - \sum_{j=1}^{n} \alpha_j M_j V_j \frac{p_{ij} S_i}{\sum_{k=1}^{n} p_{kj} N_k} - \sum_{j=1}^{n} \beta_j \sum_{k=1}^{n} p_{kj} N_k I_k \frac{p_{ij} S_i}{\sum_{k=1}^{n} p_{kj} N_k} \\
\frac{dI_i}{dt} &= -(\gamma_i + \mu_i) I_i + \sum_{j=1}^{n} \alpha_j M_j V_j \frac{p_{ij} S_i}{\sum_{k=1}^{n} p_{kj} N_k} + \sum_{j=1}^{n} \beta_j \sum_{k=1}^{n} p_{kj} N_k I_k \frac{p_{ij} S_i}{\sum_{k=1}^{n} p_{kj} N_k} \\
\frac{dR_i}{dt} &= \gamma_i I_i - \mu_i R_i \\
\frac{dU_i}{dt} &= \nu_i(1 - U_i) - \vartheta_i \frac{U_i}{M_i}\sum_{k=1}^{n} p_{ki} N_k I_k \\
\frac{dV_i}{dt} &= -\nu_i V_i + \vartheta_i \frac{U_i}{M_i}\sum_{k=1}^{n} p_{ki} N_k I_k .
\end{aligned}
$$

4 Comparison of Both Models in Three-Patch Scenario

In a case where $n = 3$ we numerically simulated both the models and compared the results. We obtained the influence of the residence time budgeting matrix on the multi-patch model. Here we restrict ourselves to consider three patches with the same set of parameters and population sizes. The movements between these three patches are defined using the residence time budgeting matrix P. The question is how far does the dynamics deviate from the single patch case, when the movement between the patches is controlled using the p_{ij} values. We use the parameters and population sizes, as given in Table 1, for the numerical simulation. We are studying two cases—the three patches being coupled and completely decoupled respectively. For the first case

$$
P = \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ 0.5 & 0.1 & 0.4 \\ 0.3 & 0.6 & 0.1 \end{pmatrix} \tag{1}
$$

Table 1 Note that the parameters and population sizes here are chosen for display of the qualitative behaviour and are not taken from any reference

μ	α	β	ϑ	ν	N	M
10/(1000*365)	0.008	0.01	0.4	1/14	20000	100000

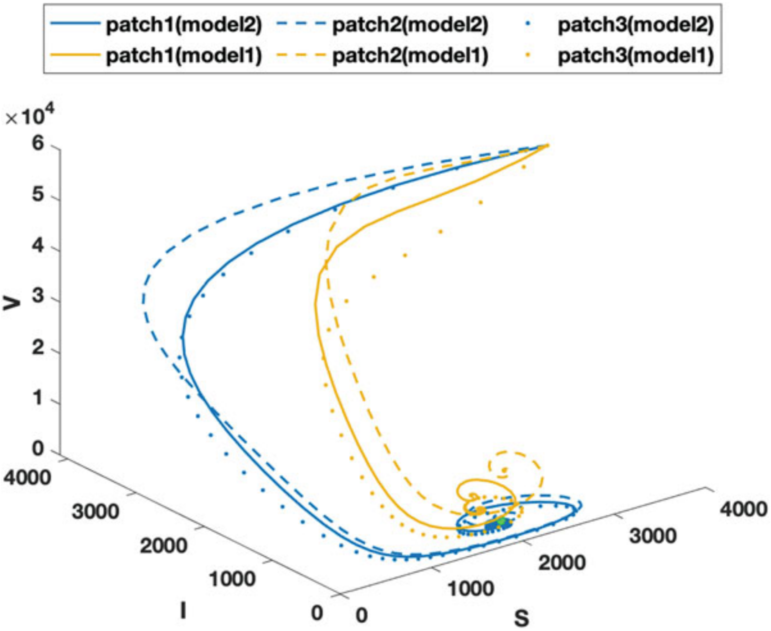


Fig. 1 Phase portrait for three patches using model 1 (Sect. 2) and model 2 (Sect. 3) for the case where the patches are coupled using the matrix P from (1)

Qualitatively, the two models exhibited similar dynamics in the case when $P \neq I$ in Fig. 1, but the results were quantitatively non-identical. The dynamics was supposed to be similar for the single patch and multi-patch models for the case $P = I$. But we have not seen this property for model 1. This is implied in Fig. 2.

5 Conclusion

In this study we have considered two different models to describe the dynamics of ZIKV spread. We compared the two models to identify the suitable model. When the commuting between patches is ceased we expect that all the three patches follow the dynamics of the single patch model. The first model failed to satisfy this condition where as the second model was successfully exhibiting this property. This gives rise to a more thorough study of the second model in a forthcoming work.

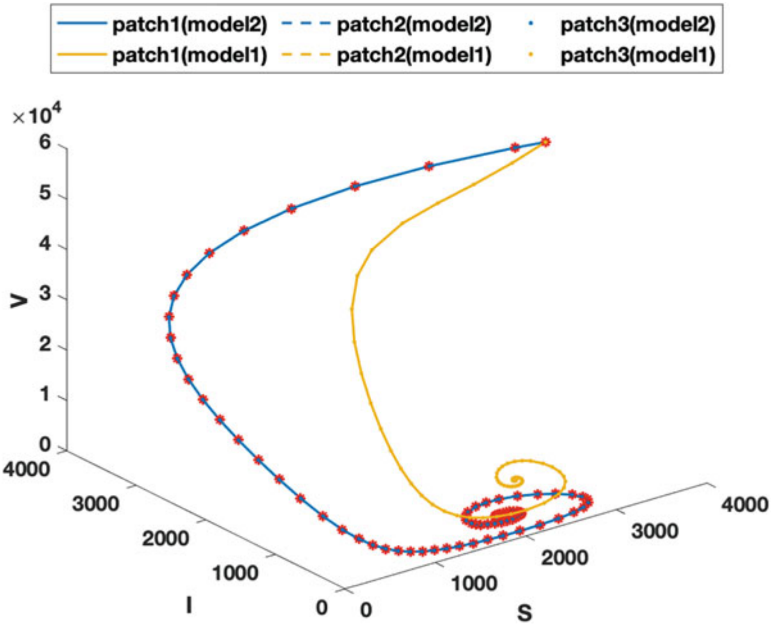


Fig. 2 For the same set of parameters as in Fig. 1 when P is set to identity matrix we see the given results where the red starred curve is the phase portrait of the single patch model

References

1. Kauffman, E. B. and Kramer, L. D.: Zika Virus Mosquito Vectors: Competence, Biology, and Vector Control. The Journal of Infectious Diseases. **216**, S976–S990 (2017)
2. Mead, P.S., Hills, S.L. and Brooks, J.T.: Zika virus as a sexually transmitted pathogen. Current Opinion in Infectious Diseases. **31** (2018)
3. Barbi, L., Coelho, A.V.C., A.D. Alencar L.C. and Crovella, S.: Prevalence of Guillain-Barré syndrome among Zika virus infected cases: a systematic review and meta-analysis. The Brazilian Journal of Infectious Diseases. **22**, 137–141(2018).
4. Araujo, A.Q.C., Silva, M.T.T. and Araujo, A.P. Q.C.: Zika virus-associated neurological disorders: a review. Brain. **139**, 2122–2130(2016)
5. Fauci, A. S. and Morens, D. M.: Zika virus in the Americas – yet another arbovirus threat. New England Journal of Medicine. **374**, 601–604(2016)
6. Heidrich, P., Jayathunga, Y., Bock W. and Götz, T.: Prediction of Dengue cases based on human mobility and seasonality – an example for the city of Jakarta. Mathematical Methods in the Applied Sciences. **44**, 13633–13658(2021)
7. Bock W. and Jayathunga, Y.: Optimal control and basic reproduction numbers for a compartmental spatial multipatch dengue model. Math. Meth. Appl. Sci. **41**, 3231–3245(2018)
8. Bichara D. and Iggidr, A.: Multi-patch and multi-group epidemic models: a new framework. Journal of Mathematical Biology. **77**, 107–134(2018)
9. Hethcote H.W.: The mathematics of infectious diseases. SIAM Review. **42**, 599–653(2000)

Discrete Port-Hamiltonian Coupled Heat Transfer

Jens Jäschke, Matthias Ehrhardt, Michael Günther, and Birgit Jacob

Abstract Heat transfer and cooling solutions play an important role in the design of gas turbine blades. However, the underlying mathematical coupling structures have not been thoroughly investigated. In a previous work, we successfully modelled a simplified version of this problem as an infinite-dimensional system. Here, we construct a spatial discretization for the above problem and investigate its properties. We show that the discrete system is less restrictive than the original infinite-dimensional system, suggesting something like a regularization effect due to discretization.

1 Introduction

The heat transfer within the blade of a gas turbine defines an important task within the simulation of gas turbines [1]. Here, we consider a simplified model system [4], where the metal of the turbine blade itself is reduced to a one-dimensional rod ($a < x_m < b$). One end of the rod is in contact with an external thermal reservoir representing the hot air driving the turbine, and the other end is in contact with the relatively cooler air flowing through the blade's cooling channel ($i < x_c < o$).

The heat transfer along the rod is modelled as a simple heat equation (index 'm') with Robin boundary conditions (also known as convective boundary conditions). The cooling channels themselves are modelled as simple transport equations, divided into an incoming channel part (index 'in') and an outgoing channel part (index 'out'), both connected to the rod at the coupling point.

Overall, we get a multiphysics model described by three coupled PDE models for the heat equation in the metal and the transport equations for the incoming and outgoing cooling air:

J. Jäschke (✉) · M. Ehrhardt · M. Günther · B. Jacob
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: jaeschke@uni-wuppertal.de; ehrhardt@uni-wuppertal.de; guenther@uni-wuppertal.de; bjacob@uni-wuppertal.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_57

Heat equation of metal

$$\frac{\partial \vartheta_m}{\partial t} = \frac{k}{c_m}\frac{\partial^2 \vartheta_m}{\partial x_m^2}, \quad a < x_m < b, \quad t > 0, \tag{1a}$$

$$-k\,\frac{\partial \vartheta_m}{\partial x_m}(a,t) = \alpha_a\big(T_{\text{ext}}(t) - \vartheta_m(a,t)\big), \quad t > 0, \tag{1b}$$

$$-k\,\frac{\partial \vartheta_m}{\partial x_m}(b,t) = \alpha_b\big(\vartheta_m(b,t) - \vartheta_{\text{in}}(c,t)\big), \quad t > 0, \tag{1c}$$

Transport of incoming cooling air

$$\frac{\partial \vartheta_{\text{in}}}{\partial t} = -v\frac{\partial \vartheta_{\text{in}}}{\partial x_c}, \quad i < x_c < c, \quad t > 0, \tag{2a}$$

$$\vartheta_{\text{in}}(i,t) = T_{\text{inlet}}(t), \quad t > 0, \tag{2b}$$

Transport of outgoing cooling air

$$\frac{\partial \vartheta_{\text{out}}}{\partial t} = -v\frac{\partial \vartheta_{\text{out}}}{\partial x_c}, \quad c < x_c < o, \quad t > 0, \tag{3a}$$

$$c_c v\big(\vartheta_{\text{out}}(c,t) - \vartheta_{\text{in}}(c,t)\big) = \alpha_b\big(\vartheta_m(b,t) - \vartheta_{\text{in}}(c,t)\big), \quad t > 0. \tag{3b}$$

In a previous work [4] we have shown that this multiphysics system can be formulated as an infinite-dimensional Port-Hamiltonian system (pHs) [2, 3]. Here, we will show that discretizing the three subsystems separately will define three index-0 port-Hamiltonian descriptor (pHDAE) systems [6] (E is the identity), which can be combined to form a single pHDAE system when properly coupled. pHDAE systems generalize the PHS setting from ODEs to DAEs. For an ease of reference we recall.

Definition 1 (Port-Hamiltonian Descriptor System, pHDAE [6]) Let $\mathcal{X} \subset \mathbb{R}^n$ the state space, $x(t) \in \mathcal{X}$ the state, $u(t),\ y(t) \in \mathbb{R}^m$ the input and output, $E \in \mathbb{R}^{l\times n}$ the flow matrix, $z \in \mathbb{R}^l$ the efforts, $J, R \in \mathbb{R}^{l\times l}$ the structure and dissipation matrices, $B, P \in \mathbb{R}^{l\times m}$ the port matrices and $S, N \in \mathbb{R}^{m\times m}$ the feed-through matrices. Then the system of differential (-algebraic) equations

$$E\dot{x} = (J - R)z + (B - P)u, \tag{4a}$$

$$y = (B + P)^\top z + (S - N)u, \tag{4b}$$

associated with the Hamiltonian function $H \in C^1(\mathcal{X}, \mathbb{R})$, is a *port-Hamiltonian descriptor system*, if the following properties hold:

1. The extended structure and dissipation matrices Γ, $W \in \mathbb{R}^{l+m\times l+m}$ defined as

$$\Gamma = \begin{pmatrix} J & B \\ -B^\top & N \end{pmatrix}, \quad W = \begin{pmatrix} R & P \\ P^\top & S \end{pmatrix} \tag{5}$$

satisfy $\Gamma = -\Gamma^\top$ and $W = W^\top \geq 0$, i.e. W is positive semi-definite.
2. $\frac{\partial H}{\partial x} = E^\top z$.

2 Discretization of the Heat Equation

We choose $I_m + 1$ grid points $x_0 = a, \dots, x_{I_m} = b$ and a step size $h = (b-a)/I_m$. We discretize the spatial derivative in (1a) by the standard second order difference quotient at $x_1, \dots, x_{I_m-1}$. Denoting the temperature at the grid points x_i by $T_i(t) = \vartheta(x_i, t)$, both boundary conditions (1a), (1c) can be solved for T_0 and $T_{I\,m}$. Summing up, we get with $T^{(m)} := (T_1, \dots, T_{I_m-1})^\top$

$$\dot{T}^{(m)} = \underbrace{\frac{k}{c_m h^2}\left(\frac{1}{1+\frac{h}{k}\alpha_a}\left(e_1 e_{I_m-1}^\top + e_{I_m-1}e_1^\top\right) + \text{tridiag}(1,-2,1)\right)}_{A_m :=} \underbrace{T^{(m)}}_{z :=} \tag{6}$$

$$+ \underbrace{\frac{k}{c_m h^2}\left(e_1\, e_{I_m-1}\right)}_{B :=} \underbrace{\begin{pmatrix} T_{\text{ext}} \\ \vartheta_{\text{in}}(c,t) \end{pmatrix}}_{u :=}.$$

With A_m, B, z and u defined above, and setting $J = 0$, $R = -A_m$, $P = 0$, $S = 0$, $N = 0$, we get the pHDAE structure of type (4). Condition (5), i.e. $W \geq 0$, holds as R is positive semi-definite due to the Gershgorin circle theorem for all physically meaningful (i.e. positive) parameters h, k and α_a, α_b.

3 Discretization of the Transport Equations

To discretize the transport equations (2a), (3a) with respect to space, we choose $I_c + 1$ grid points $x_0, \dots, x_{I_c}$ and a first-order upwind discretization (for $v \geq 0$). Replacing T_0 by the inlet boundary condition (2b), we arrive at the following semi-discrete system with $T^{(\text{in})} := (T_1, \dots, T_{I_c})^\top$:

$$\dot{T}^{(\mathrm{in})} = \underbrace{-\frac{v}{h}\,\mathrm{tridiag}(-1,1,0)}_{A_c :=}\underbrace{T^{(\mathrm{in})}}_{z :=} + \frac{v}{h}e_1 \cdot T_{\mathrm{inlet}}. \tag{7}$$

In order to get a pHDAE structure, we split the matrix of (7) into $J = \frac{1}{2}(A_c - A_c^\top)$ and $R = -\frac{1}{2}(A_c + A_c^\top)$, cf. (4) and set

$$B^\top = \frac{v}{h}\begin{pmatrix}\frac{1}{2} & 0 & \dots & 0 & \frac{1}{2}\end{pmatrix}, \quad P^\top = \frac{v}{h}\begin{pmatrix}-\frac{1}{2} & 0 & \dots & 0 & \frac{1}{2}\end{pmatrix}, \quad S = \kappa, \quad N = 0, \quad u = T_{\mathrm{inlet}},$$

with $\kappa \geq 1$. With these choices, we get

$$W = \begin{pmatrix} R & P \\ P^\top & S \end{pmatrix} = \frac{v}{h}\begin{pmatrix} \mathrm{tridiag}(-\frac{1}{2}, 1, -\frac{1}{2}) & \frac{1}{2}(-e_1 + e_{I_c}) \\ \frac{1}{2}(-e_1 + e_{I_c})^\top & \kappa \end{pmatrix}.$$

Again, the Gershgorin circle theorem yields the positive semi-definiteness of W.

For the outgoing cooling air (3a) we proceed analogously, but replace the coupling condition (3b) with a simple input similar to Eq. (2b). Equation (3b) is later included as a coupling condition in the coupled system in Sect. 4. We then arrive at the semi-discrete system

$$\dot{T}^{(\mathrm{out})} = -\frac{v}{h}\,\mathrm{tridiag}(-1,1,0)T^{(\mathrm{out})} + \frac{v}{h}e_1 \cdot T^{(\mathrm{out})}_{\mathrm{inlet}}, \tag{8}$$

with $T^{(\mathrm{out})} := (T_1, \dots, T_{I_c})^\top$. Making the same choices as above, it is obvious that this is also a pHDAE.

4 The Coupled Discrete System

In the previous sections we have formulated the semi-discretized subsystems as three port-Hamiltonian systems of the type (with $x \in \{\mathrm{m}, \mathrm{in}, \mathrm{out}\}$):

$$\dot{T}^{(x)} = (J^{(x)} - R^{(x)})T^{(x)} + (B^{(x)} - P^{(x)})u^{(x)},$$
$$y^{(x)} = (B^{(x)} + P^{(x)})^\top T^{(x)} + (S^{(x)} - N^{(x)})u^{(x)}.$$

According to [6], an interconnection of port-Hamiltonian descriptor systems (pHDAEs) (see Definition 1) is again a pHDAE if we can find an interconnection satisfying

$$Mu + Ny = 0, \tag{9}$$

with any matrices M and N. Note, however, that this does not reduce the number of inputs and outputs in general. The resulting pHDAE then has the form, cf. (5)

$$\begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{T} \\ \dot{\hat{u}} \\ \dot{\hat{y}} \end{pmatrix} = \begin{pmatrix} \Gamma - W & & 0 & 0 \\ & & I & -M^\top \\ 0 & -I & 0 & -N^\top \\ 0 & M & N & 0 \end{pmatrix} \begin{pmatrix} T \\ \hat{u} \\ \hat{y} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ I \\ 0 \end{pmatrix} u, \tag{10}$$

$$y = \hat{y}, \tag{11}$$

with

$$T(t) = \begin{pmatrix} T^{(m)}(t) \\ T^{(\text{in})}(t) \\ T^{(\text{out})}(t) \end{pmatrix} \in \mathbb{R}^{I_m - 1 + 2I_c}, \qquad \hat{u}(t), \hat{y}(t) \in \mathbb{R}^4,$$

$$\Gamma - W = \Pi \ \operatorname{diag}\left(\Gamma^{(m)} - W^{(m)}, \Gamma^{(\text{in})} - W^{(\text{in})}, \Gamma^{(\text{out})} - W^{(\text{out})}\right)\Pi^\top,$$

as in Definition 1 with a permutation matrix

$$\Pi = \begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}.$$

It is worth mentioning that the additionally introduced variables $\hat{u}$ and $\hat{y}$ are just copies of the inputs u and outputs y. Note that the above property makes no statement about the index of the resulting pHDAE. While this is common also for coupling "regular" ODEs, it is important to keep in mind, since even when all subsystems are index-0 (i.e. ODEs), the coupled system can have a higher index.

We can now check whether the coupled system (10) exhibits the form (9). The inputs u and outputs y of the coupled system (10) are

$$\text{inputs:} \quad u = \begin{pmatrix} u_1^{(m)} \\ u_2^{(m)} \\ u^{(\text{in})} \\ u^{(\text{out})} \end{pmatrix} = \begin{pmatrix} T_{\text{ext}}^{(m)} \\ \vartheta_{\text{in}}(c, t) \\ T_{\text{inlet}}^{(\text{in})} \\ T_{\text{inlet}}^{(\text{out})} \end{pmatrix}, \tag{12}$$

$$\text{outputs:} \quad y = \begin{pmatrix} y_1^{(m)} \\ y_2^{(m)} \\ y^{(\text{in})} \\ y^{(\text{out})} \end{pmatrix} = \begin{pmatrix} \frac{k}{c_m h^2} \frac{1}{1+\frac{1}{\frac{h}{k}\alpha_a}} T_1^{(m)} \\ \frac{k}{c_m h^2} \frac{1}{1+\frac{1}{\frac{h}{k}\alpha_b}} T_{I_m-1}^{(m)} \\ \frac{v}{h} T_{I_c}^{(\text{in})} + \kappa T_{\text{inlet}}^{(\text{in})} \\ \frac{v}{h} T_{I_c}^{(\text{out})} + \kappa T_{\text{inlet}}^{(\text{out})} \end{pmatrix}, \qquad y^{(m)} = \begin{pmatrix} y_1^{(m)} \\ y_2^{(m)} \end{pmatrix}. \tag{13}$$

The input of the heat equation still references $\vartheta_{\text{in}}(c, t)$, a quantity of the continuous system. From Eq. (2a) as well as Eq. (7), we can see that it is equivalent to $T_{I_c}^{(\text{in})}$ of the discrete cooling channel:

$$u_2^{(m)} = \vartheta_{\text{in}}(c, t) = T_{I_c}^{(\text{in})} = \frac{h}{v} y^{(\text{in})} - \frac{h\kappa}{v} u^{(\text{in})}.$$

Equation (3b) yields the coupling condition

$$c_c v T_0^{(\text{out})} - c_c v T_{I_c}^{(\text{in})} = \alpha_b T_{I_m}^{(m)} - \alpha_b T_{I_c}^{(\text{in})}$$

to the outgoing cooling channel, i.e. using (12), (13) and the explicit formula of T_{I_m}

$$c_c v u^{(\text{out})} - c_c v \frac{h}{v}\left(y^{(\text{in})} - \kappa u^{(\text{in})}\right) = c_m h y_2^{(m)} + \frac{\alpha_b}{1 + \frac{1}{\frac{h}{k}\alpha_b}} u_2^{(m)} - \alpha_b \frac{h}{v}\left(y^{(\text{in})} - \kappa u^{(\text{in})}\right).$$

Together this leads to an interconnection relation of the form (9)

$$\underbrace{\begin{pmatrix} 0 & 1 & \frac{h\kappa}{v} & 0 \\ 0 & -\frac{\alpha_b}{1+\frac{1}{\frac{h}{k}\alpha_b}} & (c_c h - \frac{h}{v}\alpha_b)\kappa & c_c v \end{pmatrix}}_{M} u + \underbrace{\begin{pmatrix} 0 & 0 & -\frac{h}{v} & 0 \\ 0 & -c_m h & \alpha_b \frac{h}{v} - c_c h & 0 \end{pmatrix}}_{N} y = 0,$$

and therefore, the considered coupled system is a pHDAE.

However, the above model does not define a Dirac structure for (y, u) and is therefore not an energy-conserving coupling in terms of the quantity acting as energy in the Hamiltonian under consideration, i.e., not the physical energy. Thus, the criteria [6] for index reduction and row operations to reduce the system are not satisfied.

5 Conclusion

We have found that the multiphysics approach to discretization before coupling works quite well and requires only a small change in our transport equations. Interestingly, unlike the continuous system, it has no constraints on the parameters, but leads to a pHDAE that potentially has a nonzero index. In future work, following the ideas of Kotyczka and Lefèvre [5], we will consider our multiphysics problem as a discrete-time port Hamiltonian system arising from a discrete-time Dirac structure, that is obtained by a symplectic Gauss-Legendre collocation method.

References

1. Backhaus, J., Bolten, M., Doganay, O.T., et al.: GivEn – Shape optimization for gas turbines in volatile energy networks. In: S. Göttlich, M. Herty, A. Milde (eds.) Mathematical MSO for Power Engineering and Management, Math. Industry, vol. 34, pp. 71–106. Springer (2021)
2. Beattie, C., Mehrmann, V., Xu, H., Zwart, H.: Linear port-Hamiltonian descriptor systems. Mathematics of Control, Signals, and Systems **30**(4), 17 (2018).
3. Jacob, B., Zwart, H.J.: Linear port-Hamiltonian systems on infinite-dimensional spaces, vol. 223. Springer Science & Business Media (2012)
4. Jäschke, J., Ehrhardt, M., Günther, M., Jacob, B.: A port-Hamiltonian formulation of coupled heat transfer. Mathematical and Computer Modelling of Dynamical Systems, 2022.
5. Kotyczka, P., Lefèvre, L.: Discrete-time port-Hamiltonian systems based on Gauss-Legendre collocation. IFAC-PapersOnLine **51**(3), 125–130 (2018)
6. Mehrmann, V., Morandin, R.: Structure-preserving discretization for port-Hamiltonian descriptor systems. In: 2019 IEEE 58th Conf. Decision and Control (CDC), pp. 6863–6868 (2019).

A Non-Reflecting Boundary Condition for Multispeed Lattice Boltzmann Methods

Friedemann Klass, Alessandro Gabbana, and Andreas Bartel

Abstract Artificial boundary conditions are commonly employed in numerical simulations to confine very large or unbounded domains to a computationally feasible finite domain. The implementation of an artificial boundary condition should cause no interaction with the bulk dynamics, and in particular should not create artifacts such as reflections of pressure waves. In the context of the Lattice Boltzmann Method (LBM), standard velocity or pressure boundary conditions do not fulfill this requirement. This problem is further emphasized when using multispeed LBM models, in which several layers of boundary nodes interact with the bulk dynamics. In this work, we take a first step towards the definition of a discrete artificial boundary condition for LBM based on stencils with multiple speed levels.

1 Introduction

Artificial boundaries are commonly posed whenever the domain of interest is embedded in a large or even unbounded domain, e.g. in numerical studies in the field of astrophysics or acoustics. Their key property is to have no effect on the bulk dynamics, since they should only serve in the truncation of the computational domain. In the mesoscopic framework of the Lattice Boltzmann Method (LBM), where Boundary Conditions (BC) are set in terms of values assigned to discrete distribution functions [11], this task is more complex than in the macroscopic case. Approaches to model artificial boundaries in the LBM found in the literature include characteristic BC [3], absorbing layers [7] and the so-called Discrete

F. Klass (✉) · A. Bartel
University of Wuppertal, Wuppertal, Germany
e-mail: fklass@uni-wuppertal.de; bartel@uni-wuppertal.de

A. Gabbana
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: a.gabbana@tue.nl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_58

Artificial BC (DABC) [1, 2]. Implementation of the latter requires the solution of a subproblem that uses information from previous time steps to mimic the systems evolution on a larger grid. High order LBM based on multispeed lattices represent a powerful framework which allows the extension of the applicability, potentially beyond Navier-Stokes [8, 9, 12]. However, they are not commonly adopted since the definition of suitable BC is made difficult by the presence of multiple speed levels. Few BCs for such multispeed (thermal) lattices can be found in the literature, e.g. [4, 6] and, to the best of our knowledge, there are currently no explicit formulations of artificial BC for multispeed lattices. In this work, we take a first step towards the definition of accurate and efficient DABC for multispeed LBM.

This article is organized as follows: In Sect. 2, we summarize the LBM algorithm. In Sect. 3, we describe the DABC and discuss its extension to multispeed lattices. We validate this extension on numerical results in Sect. 4 before summarizing our findings in Sect. 5.

2 The Lattice Boltzmann Method (LBM)

LBM [11] is a mesoscopic fluid dynamics solver, discrete in time and velocity, where the description of a fluid in d space dimensions is based on the synthetic dynamics of a set of populations f_i sitting at discrete lattice sites. The discretization of the velocity space is typically coupled to a Gauss-Hermite quadrature, which ensures that all the moments of the equilibrium distribution function are exactly preserved up to a certain desired order. The q abscissae of the quadrature, $\mathbf{c}_1, \dots, \mathbf{c}_q$, form the velocity stencil, which dictates how information propagates at each time step. Depending on the specific stencil adopted, it is common to distinguish between the different LBM models using the notation DdQq. Here we consider the D2Q17 model, Fig. 1, which is a 7th-order quadrature rule [8, 10]. Quadrature weights are given in Table 1.

The evolution of the system is governed by the lattice Boltzmann equation:

$$f_i(\mathbf{x} + \mathbf{c}_i \Delta t, t + \Delta t) = f_i(\mathbf{x}, t) - \tfrac{\Delta t}{\tau}\big(f_i(\mathbf{x}, t) - f_i^{\mathrm{eq}}(\mathbf{x}, t)\big) \tag{1}$$

with time step Δt, and relaxation rate τ towards the discrete analogon of the Maxwell-Boltzmann distribution (using macroscopic density ρ and velocity $\mathbf{u}$)

$$f_i^{\mathrm{eq}}(\rho, \mathbf{u}) = w_i \rho \left(1 + \mathbf{u} \cdot \mathbf{c}_i + \tfrac{1}{2c_s^2}\left((\mathbf{u} \cdot \mathbf{c}_i)^2 - u^2\right) + \tfrac{\mathbf{u} \cdot \mathbf{c}_i}{6c_s^4}\left((\mathbf{u} \cdot \mathbf{c}_i)^2 - 3u^2\right)\right) \tag{2}$$

with lattice speed of sound c_s and the quadrature weights w_i (Table 1).

The LBM algorithm is based on the so-called stream and collide paradigm: The right hand side of Eq. (1) is referred to as collision step, driving populations towards the local equilibrium, while the left hand side of Eq. (1) corresponds to the

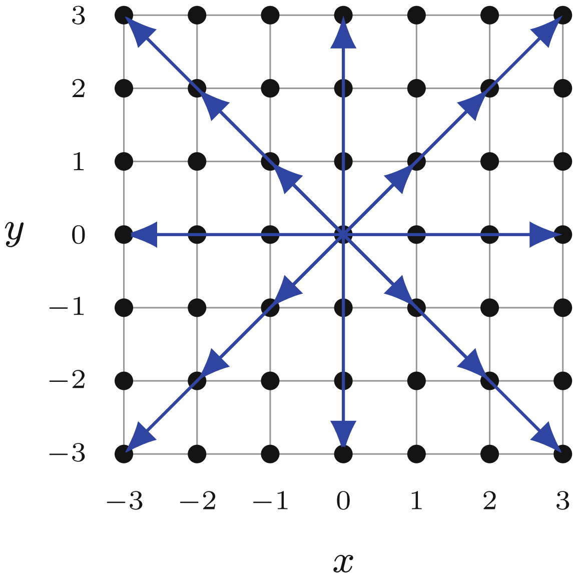


Fig. 1 Velocities $\mathbf{c}_i$ for the D2Q17 stencil. It is a minimal set for square lattices to *exactly* recover the moments of the distribution up to the third order [10]

Table 1 D2Q17 quadrature weights w_i for each velocity group. FS means full-symmetric, i.e., $(\pm 1, \pm 1)_{\mathrm{FS}} = \{(-1, -1), (-1, 1), (1, -1), (1, 1)\}$, and c_s is the lattice speed of sound

$\mathbf{c}_i$	w_i
$(\ 0,\ \ 0)$	0.121527777777778
$(\pm 1, \pm 1)_{\mathrm{FS}}$	0.175781250000000
$(\pm 2, \pm 2)_{\mathrm{FS}}$	0.014062500000000
$(\pm 3, \pm 3)_{\mathrm{FS}}$	0.001996527777778
$(\ 0, \pm 3)_{\mathrm{FS}}$	0.027777777777778
c_s	1.224744871391589

streaming step, in which populations are assigned to neighbouring nodes following the directions $\mathbf{c}_i$. Macroscopic density ρ and velocity $\mathbf{u}$ are defined as the velocity moments of the particle distribution. Thanks to the underlying quadrature rule, they can be computed as summations over the discrete populations:

$$\rho = \sum_{i=1}^{q} f_i, \quad \rho\mathbf{u} = \sum_{i=1}^{q} f_i \mathbf{c}_i. \tag{3}$$

3 Discrete Artificial Boundary Condition

The task of any Boundary Condition (BC) in the LBM context is to set the populations that remain unspecified after the streaming step. Standard BCs will impose a macroscopic pressure or velocity. Such a constraint leads to a system of equations in the unknown populations. However, this will give rise to reflections which are propagating into the bulk of the computational domain. Since an artificial boundary just accounts for the need of truncating the physical domain, this behaviour is clearly unphysical. Thus, artificial boundaries should be treated with a non-reflecting BC.

In the Discrete Artificial Boundary Condition (DABC) [1], the unknown populations are obtained by solving a so-called subproblem, a separate LBM simulation that takes into account information from a specified number of preceding iterations/time steps. This strategy has the advantage of working exactly in the same mesoscopic framework as the original LBM simulation and no additional structural assumptions have to be made.

Let us revise the general procedure [2], by considering a rectangular computational grid of size $L_x \times L_y$ and a maximal history depth H_{max}. For the DABC on the right boundary, the computational domain of the subproblem is $(H + 1) \times L_y$, where the history depth $H = \min(\text{iter}, H_{\text{max}})$ is the number of previous time steps to be taken into account.

By interpreting the subproblem as an extension of the original grid by H layers, we can classify any node as either belonging to the original grid, the subproblem, or their intersection Γ. That is, Γ is the set of nodes that form the right boundary of the original problem and the left boundary of the subproblem.

Algorithm 1 Right LBM subproblem for time level t_k

```
Inputs:
    max. $M$, grid size $(M(H+1), ny)$, $\tau$, max. $H$, initial fields: $\rho, \mathbf{u}$
Initialize:
    Set $f^{\text{sub}} = f^{\text{eq}} = f^{\text{eq}}(\rho, \mathbf{u})$, $\text{iter}_{\text{sub}} = 1$ and $f^{\text{sub}}(x_\gamma, t_{k-H}) = f(x_\gamma, t_{k-H})$
while $\text{iter}_{\text{sub}} \leq$ H do
    for all Gridpoints do
        Update equilibrium distribution $f^{\text{eq}}$
        Collide & Stream
        if $\text{iter}_{\text{sub}} <$ H then
            Left BC of subproblem: $f_i^{\text{sub}}(x_\gamma, t_{k-H+\text{iter}_{\text{sub}}}) = f_i(x_\gamma, t_{k-H+\text{iter}_{\text{sub}}})$
        else
            Right BC of original problem: $f_i(x_\gamma, t_k) = f_i^{\text{sub}}(x_\gamma, t_k)$
        end if
        Update macroscopic quantities
    end for
    $\text{iter}_{\text{sub}} = \text{iter}_{\text{sub}} + 1$
end while
```

Now, let us assume we are at time t_k. The nodes $x_\gamma \in \Gamma$ are initialized with populations from time t_{k-H}, while the remaining nodes of the subproblem are set to a given equilibrium. Then, we proceed with the usual LBM scheme, using

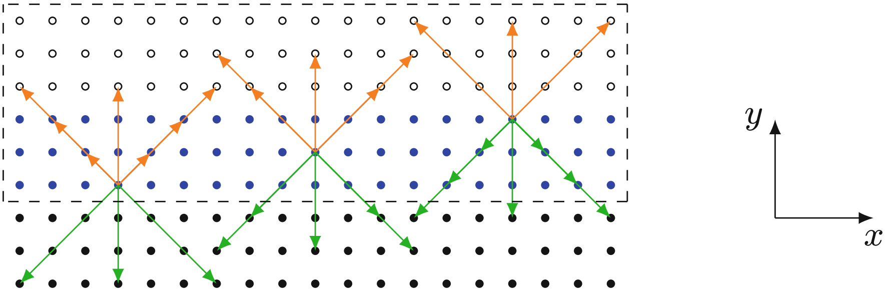


Fig. 2 Sketch of populations assigned in a subproblem for the lower boundary. The dashed rectangle contains the subproblems domain. Black and hollow nodes are fluid nodes for the original problem and subproblem, resp. Intersection Γ is colored in blue. At each iteration in the subproblem, orange populations are taken as a BC from the history of the original lattice at time $t_{k-H+\text{iter}}$. Green populations are the final output of the subproblem at time level t_k

the previously computed populations $f_i(x_\gamma,\ t_{k-H+\text{iter}})$ as a BC on Γ. Finally, the unknown populations of the original problem are obtained as post-streaming populations of the subproblem at time t_k. We remark that (i) there is no need for a BC at the right boundary of the subproblem, since any error will not propagate to the original grid and (ii) the given initial equilibrium should encode any external information available to obtain accurate results. See [1] for a discussion on initialization strategies.

When considering multispeed stencils, two extra ingredients need to be taken into account. First, the amount of layers in the subproblem has to be multiplied by the maximal horizontal displacement M of the stencil, since populations can propagate to their M-th nearest neighbours. Second, Γ will consist of M layers of nodes. Both aspects are depicted in Fig. 2 (for lower boundary), while the procedure for solving a right subproblem (i.e., an east artificial boundary) is summarized in Algorithm 1.

4 Numerical Results

We consider a simple isothermal flow created by an isolated vortex, which travels towards the right boundary ($x = 0.75$) in the spatial domain $[-0.75, 0.75]\times[-3, 3]$. The space is discretised with step size $h = 0.01$. The initial fields are assigned as:

$$u(x, y) = u_0 + \begin{cases} 0 & \text{if } x^2 + y^2 \geq r^2 \\ v(x, y) & \text{otherwise} \end{cases}, \quad v(x, y) = \frac{1}{2} 2^{-\frac{x^2+y^2}{b^2}} \begin{pmatrix} y \\ -x \end{pmatrix} c_s,$$

with $u_0 = c_s \cdot (0.2,\ 0)^T$, $r = 0.7$, $b = 0.15$ and $\rho(x, y) = 1$ and the subproblems are initialized with a density of unity and velocity of u_0. The simulation is conducted at a fixed Reynolds number Re $= 10$. We measure the L^2-errors in the macroscopic

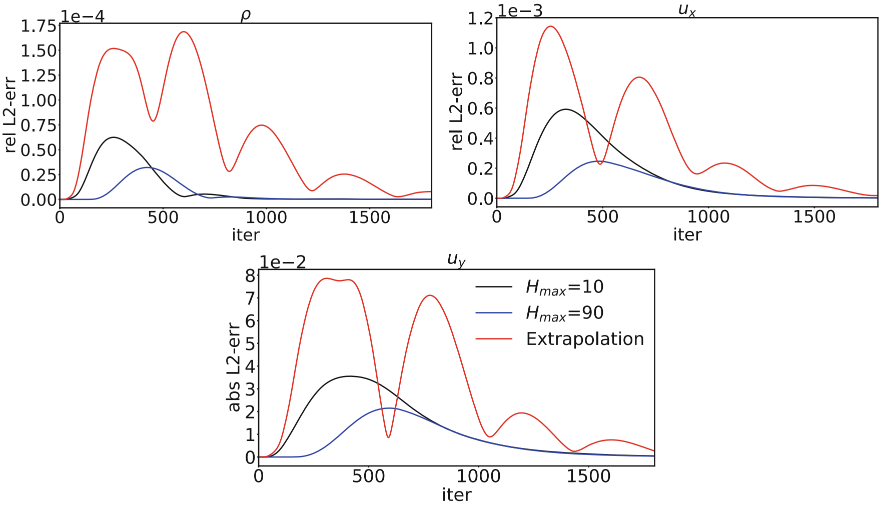


Fig. 3 Evolution of L2-errors in density and velocity. The DABC at various values for $H_{\max}$ is compared to the extrapolation BC. Top panels: Relative errors in ρ and u_x. Lower Panel: Absolute error in u_y. The D2Q17 stencil was used for simulation

fields with respect to reference fields $\rho^{\text{ref}}, \mathbf{u}^{\text{ref}}$. Such fields are obtained from a reference LBM simulation that uses a sufficiently large computational domain, ensuring that no information from the boundaries propagates into the domain of interest for the relevant amount of iterations. To simplify our analysis, the effect of the left boundary is neglected by imposing the corresponding populations from the reference simulation at the appropriate time. Upper and lower boundaries are periodic. The right boundary is equipped with the DABC.

For comparison, we also equip the right boundary with a second order extrapolation scheme [5], where the unknown populations at the right boundary are defined as

$$f_i(x_i, y_j) = \frac{4f_i(x_{i-1}, y_j) - f_i(x_{i-2}, y_j)}{3}.$$

In Fig. 3 we show the time evolution of the error for the different macroscopic fields, comparing the results obtained with the DABC with different $H_{\max}$ with those given by the extrapolated BC. We observe that the extrapolated BC causes significantly larger reflections, causing the error in the macroscopic fields to oscillate with decaying amplitude. The DABC does not exhibit this behaviour. Instead, the error initially grows as the vortex starts to interact with the boundary but then quickly drops, asymptotically approaching zero. As expected, usage of a higher $H_{\max}$ leads to lower errors. Since the reference value of u_y is zero, we show the absolute error in u_y in the lower panel of Fig. 3.

5 Conclusion

In this work we have extended a DABC to the D2Q17 LBM. Our numerical results have shown that non-reflecting BC provide significantly higher accuracy over standard BC for the modeling of artificial boundaries. We consider this to be a promising first step towards the development of DABC for multispeed LBM, capable of combining accuracy and computational efficiency. In particular the latter aspect will be object of future studies, where we will analyze the thread-off between performances and accuracy while varying $H_{\max}$. Finally, a more thorough study of initialization strategies for the subproblems and a comparison with other approaches, like the perfectly matched layer BC, appear promising candidates for future research.

References

1. D. Heubes, *Artificial Boundary Conditions in the Lattice Boltzmann Method*, PhD Dissertation, University of Wuppertal, 2017.
2. D. Heubes, A. Bartel, M. Ehrhardt, *Discrete Artificial Boundary Conditions for the Lattice Boltzmann Method in 2D*, ESAIM: Proceedings and Surveys, 2015, 47–65.
3. D. Heubes, A. Bartel, M. Ehrhardt, *Characteristic boundary conditions in the lattice Boltzmann method for fluid and gas dynamics*, J. Comput. Appl. Math. 262 (2014), 51–61.
4. F. Klass, A. Gabbana, A. Bartel, *A non-equilibrium bounce-back boundary condition for thermal multispeed LBM*, J. Comput. Sci 53 (2021), 101364.
5. H.C. Lee, S. Bawazeer, A.A. Mohamad, *Boundary conditions for lattice Boltzmann method with multispeed lattices*, Comput Fluids 162 (2018), 152–159.
6. J. Meng, Y. Zhang, *Diffuse reflection boundary condition for high-order lattice Boltzmann models with streaming-collision mechanism*, J. Comput. Phys 258 (2014), 601–612.
7. A. Najafi-Yazdi, L. Mongeau, *An absorbing boundary condition for the lattice Boltzmann method based on the perfectly matched layer*, Comput Fluids 68 (2012), 203–218.
8. P.C. Philippi, L.A. Hegele Jr., L.O.E. dos Santos, R. Surmas, *From the continuous to the lattice Boltzmann equation: The discretization problem and thermal models* Phys. Rev. E 73 (2006), 056702.
9. X. Shan, X.F. Yuan, H. Chen, *Kinetic theory representation of hydrodynamics: a way beyond the Navier–Stokes equation*, J. Fluid Mech. 550 (2006), 413–441.
10. X. Shan, *General solution of lattices for Cartesian lattice Bhatanagar-Gross-Krook models*, Phys. Rev. E 81 (2010), 036702.
11. S. Succi, *The Lattice Boltzmann Equation: For Complex States of Flowing Matter* (Oxford University Press, Oxford, 2018).
12. S. Succi, *Lattice Boltzmann beyond Navier-Stokes: Where do we stand?*, AIP Conference Proceedings 1786 (2016), 030001.

Correlation Matrices Driven by Stochastic Isospectral Flows

Michelle Muniz, Matthias Ehrhardt, and Michael Günther

Abstract In many important areas of finance and risk management, time-dependent correlation matrices must be specified. We create valid correlation matrices by extending the idea of correlation flows based on isospectral flows. To incorporate the stochastic behavior of correlations, we adapt this approach by modeling the isospectral flow as a stochastic differential equation (SDE) instead of an ordinary differential equation (ODE).

The solution of this SDE lies on the manifold of symmetric and positive semi-definite matrices, so structure-preserving schemes are needed for its numerical approximation. We apply stochastic Lie group methods based on Runge-Kutta–Munthe-Kaas schemes for ODEs to guarantee that the numerical solution evolves on the correct manifold. We also present an application example to illustrate our methodology.

1 Introduction

In this paper, we construct time-dependent correlation matrices that approximate the true correlation using real market data, reflect the stochastic nature of correlations, and satisfy the following properties of a valid correlation matrix:

1. All diagonal elements of a correlation matrix are equal to one and absolute values of all non-diagonal elements are less than or equal to one.
2. Correlation matrices are real symmetric and positive semi-definite, i.e. all eigenvalues are non-negative.

To ensure these properties, we take up the idea presented in [3, 6]. The authors constructed *covariance flows*, i.e., covariance matrices based on the isospectral flux

M. Muniz (✉) · M. Ehrhardt · M. Günther
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: muniz@uni-wuppertal.de; ehrhardt@uni-wuppertal.de; guenther@uni-wuppertal.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_59

$$\dot{P}_t = [Y_t, P_t] , \quad t \geq 0 , \tag{1}$$

where P_0 is a given valid covariance matrix, i.e. symmetric and positive semi-definite, Y_t is a skew-symmetric matrix, $Y_t \in \mathfrak{so}(n)$, and $[A, B] = AB - BA$ is the matrix commutator. The solution P_t is a differential curve on the manifold

$$\widehat{\mathrm{Sym}}(n) = \{P_t = Q_t P_0 Q_t^\top : Q_t \in \mathrm{SO}(n),\ P_0 \text{ positive semi-definite}\} , \tag{2}$$

where SO(n) denotes the space of orthogonal matrices with determinant +1. Note that the matrices in $\widehat{\mathrm{Sym}}(n)$ are similar to P_0.

The corresponding *correlation flow* is obtained by the transformation $R_t = \Sigma_t^{-1} P_t \Sigma_t^{-1}$ with $\Sigma_t = \big(\mathrm{diag}(P_t)\big)^{1/2}$.

Our goal is to extend this approach by incorporating the stochastic behavior of correlations. To this end, we formulate an isospectral flow based on (1) driven by a stochastic differential equation (SDE) rather than an ordinary differential equation (ODE). Since the solution of this SDE evolves on the manifold $\widehat{\mathrm{Sym}}(n)$, we need a method for its numerical approximation that preserves the geometric properties of the manifold. Therefore, we will present a structure-preserving Euler-Maruyama scheme based on Runge-Kutta-Munthe-Kaas (RKMK) schemes for ODEs on manifolds [5]. Further details on stochastic RKMK schemes can be found in [2, 4].

The remainder of the paper is organized as follows. In Sect. 2 we construct covariance flows based on an isospectral flow driven by a SDE. Since correlation matrices play an important role in finance and risk management we provide an application example of our methodology from the viewpoint of a risk manager using real market data in Sect. 3. A conclusion of our results is given in Sect. 4.

2 Covariance Flows Based on Stochastic Isospectral Flows

The space of covariance matrices $\widehat{\mathrm{Sym}}(n)$ is a homogeneous manifold, i.e. there exists an element Q in a corresponding Lie group such that $\Lambda(Q, P_1) = P_2$ for two arbitrary elements P_1 and P_2 of the manifold. The considered Lie group regarding isospectral flows is the space of rotation matrices SO(n) and the map $\Lambda \colon \mathrm{SO}(n) \times \widehat{\mathrm{Sym}}(n) \to \widehat{\mathrm{Sym}}(n)$, called the *Lie group action*, can be chosen as

$$\Lambda(Q, P) = QPQ^\top , \tag{3}$$

see [5]. Corresponding to this Lie group action there exists a *Lie algebra action* $\lambda \colon \mathfrak{so}(n) \times \widehat{\mathrm{Sym}}(n) \to \widehat{\mathrm{Sym}}(n)$ given by

$$\lambda(\Omega, P) = \exp(\Omega) P \exp(-\Omega) , \tag{4}$$

where the Lie algebra $\mathfrak{so}(n)$ is the tangent space at the identity I of the Lie group $\mathrm{SO}(n)$, i.e. $\mathfrak{so}(n) = T_I\mathrm{SO}(n)$, which is the space of skew-symmetric matrices.

The matrix exponential $\exp\colon \mathfrak{so}(n) \to \mathrm{SO}(n)$, $\Omega \mapsto \sum_{k\geq 0} \Omega^k/k!$ acts as a map from the Lie algebra to the Lie group and its derivative is given by

$$\left(\frac{\mathrm{d}}{\mathrm{d}\Omega}\exp(\Omega)\right)H = \left(\mathrm{d}\exp_\Omega(H)\right)\exp(\Omega), \quad \mathrm{d}\exp_\Omega(H) = \sum_{k\geq 0}\frac{1}{(k+1)!}\,\mathrm{ad}_\Omega^k(H)\ , \tag{5}$$

see [1, p. 83]. By $\mathrm{ad}_\Omega(H) = [\Omega, H] = \Omega H - H\Omega$ we express the adjoint operator

$$\mathrm{ad}_\Omega^0(H) = H, \quad \mathrm{ad}_\Omega^k(H) = \left[\Omega, \mathrm{ad}_\Omega^{k-1}(H)\right] = \mathrm{ad}_\Omega\left(\mathrm{ad}_\Omega^{k-1}(H)\right), \quad k \geq 1\ .$$

Theorem 1 *Assume that* $\mathrm{d}\exp_\Omega(H)$ *in* (5) *is invertible and let* $\Omega_t \in \mathfrak{so}(n)$ *be driven by*

$$\mathrm{d}\Omega_t = A_t\mathrm{d}t + \sum_{i=1}^m \Gamma_{i,t}\mathrm{d}W_{i,t}\ , \quad \Omega_0 = 0\ . \tag{6}$$

Then $P_t = \exp(\Omega_t)P_0\exp(-\Omega_t)$ *obeying*

$$\mathrm{d}P_t = \left(\left[Y_{0,t}, P_t\right] + \frac{1}{2}\sum_{i=1}^m \left[Y_{i,t}, \left[Y_{i,t}, P_t\right]\right]\right)\mathrm{d}t + \sum_{i=1}^m \left[Y_{i,t}, P_t\right]\mathrm{d}W_{i,t} \tag{7}$$

is an isospectral flow in $\widehat{\mathrm{Sym}}(n)$*, where* $Y_{i,t} \in \mathfrak{so}(n)$ *for* $i = 0, \ldots, m$.

The coefficients in (6) *are given by*

$$A_t = \mathrm{d}\exp_{\Omega_t}^{-1}\left(Y_{0,t} - \frac{1}{2}\sum_{i=1}^m C_{i,t}\right)\ , \qquad \Gamma_{i,t} = \mathrm{d}\exp_{\Omega_t}^{-1}(Y_{i,t})\ ,$$

where

$$\begin{aligned} C_{i,t} &= \left(\frac{\mathrm{d}}{\mathrm{d}\Omega}\mathrm{d}\exp_{\Omega_t}(\Gamma_{i,t})\right)\Gamma_{i,t} \\ &= \sum_{k=0}^\infty\sum_{j=0}^\infty \frac{1}{(k+j+2)}\frac{(-1)^{j+1}}{k!(j+1)!}\,\mathrm{ad}_{\Omega_t}^k\left(\mathrm{ad}_{\Gamma_{i,t}}\left(\mathrm{ad}_{\Omega_t}^j(\Gamma_{i,t})\right)\right)\ . \end{aligned}$$

The SDE (7) and the coefficients in (6) can be derived by applying Itô's lemma to $P_t = \exp(\Omega_t)P_0\exp(-\Omega_t)$ and assuming an additive perturbation by independent Wiener processes $W_{1,t}, \ldots, W_{m,t}$ to the ODE (1). Since $\exp(\Omega_t)P_0\exp(-\Omega_t)$ corresponds to the Lie algebra action (4) with $P \equiv P_0$, the solution P_t will evolve in $\widehat{\mathrm{Sym}}(n)$ by construction.

The expression $\mathrm{d}\exp_{\Omega}(H)$ in (5) is invertible if the eigenvalues of ad_{Ω} are different from $2\ell\pi i$ with $\ell \in \{\pm 1, \pm 2, \dots\}$. The inverse converges for $\|\Omega\| < \pi$ and is given by

$$d\exp_{\Omega}^{-1}(H) = \sum_{k=0}^{\infty} \frac{B_k}{k!}\, \mathrm{ad}_{\Omega}^{k}(H)\ , \tag{8}$$

where B_k denotes the Bernoulli numbers (see Lemma III.4.2 (Baker, 1905) in [1]).

Note that the assumption of a SDE in the Lie algebra gives the benefit of applying actions in a linear space whereas applying linear actions to (7) on the manifold $\widehat{\mathrm{Sym}}(n)$ would result in a *drift-off*.

3 Simulation of Correlation Flows

We assume the following scenario: A risk manager retrieves from the middle office's reporting system the initial correlation matrix

$$R_0^{\text{hist}} = \begin{pmatrix} 1 & -0.0159 \\ -0.0159 & 1 \end{pmatrix}\ , \tag{9}$$

of the moving correlations between the S&P 500 index and the Euro/US-Dollar exchange rate on a daily basis computed with a window size of 30 days from January 3, 2005 to January 6, 2006 seen in Fig. 1. Furthermore, we assume that the risk manager is aware of the density function of the considered correlation as the path shown in Fig. 1 is only one of many possible realizations. Therefore, we estimate a density function from the historical data using kernel smoothing functions (see Fig. 2). Now, the risk manager's task is to create valid time-dependent correlation matrices that reflect the stochastic nature of correlations while trying to match the density function of the historical data.

Our proposed methodology for the risk manager is given by the following steps:

1. Compute a covariance matrix P_0 based on the historical correlation matrix R_0^{hist} and consider the covariance flow $P_t = \exp(\Omega_t)P_0\exp(-\Omega_t)$ obeying (7) where the skew-symmetric matrices $Y_{0,t}, \dots, Y_{m,t}$ are set such that they contain parameters as degrees of freedom.
2. Solve the SDE (6) in the Lie algebra numerically and define a solution of (7) according to the Lie algebra action (4). Transform the obtained covariance matrices to corresponding correlation matrices.
3. Estimate the density function from the so-obtained correlation flow and calibrate the involved parameters such that the density function of the correlation flow matches the density function of the historical correlation.

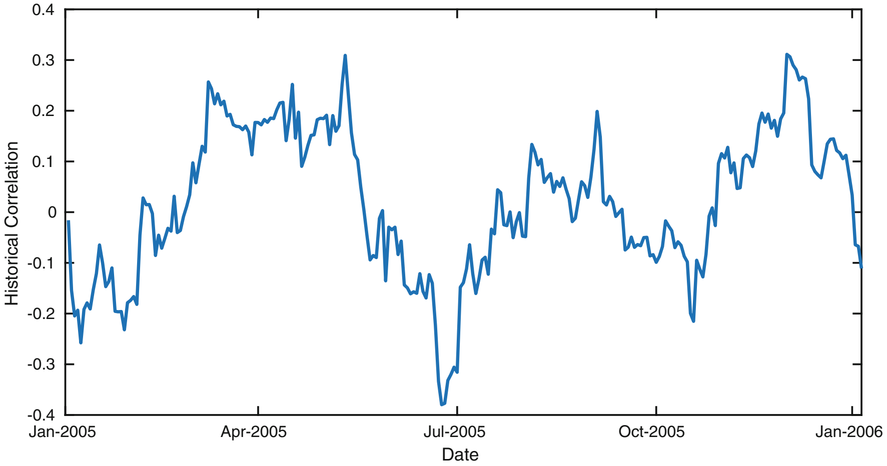


Fig. 1 The 30-day historical correlations between S&P 500 and Euro/US-Dollar exchange rate, source of data: www.yahoo.com

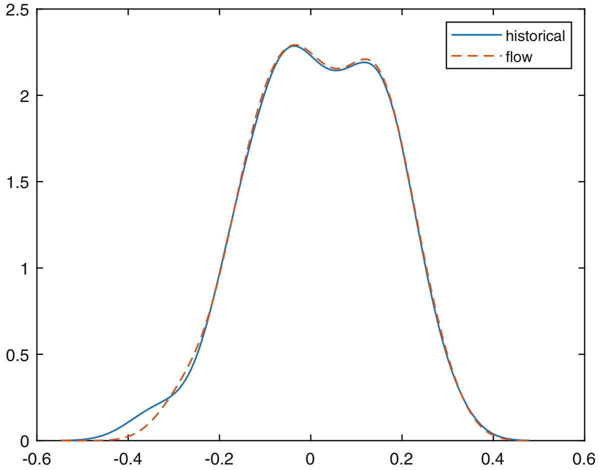


Fig. 2 Empirical density function of the historical correlation and the correlation flow between S&P 500 and Euro/US-Dollar exchange rate, computed with the MATLAB function `ksdensity`

These steps are now specified for $n = 2$ and $m = 2$.

Setting P_0 and $Y_{0,t}, Y_{1,t}, Y_{2,t}$

For the construction of P_0 we set D as the diagonal matrix containing the eigenvalues of the estimated covariance matrix of the whole historical data and we tried to find an orthogonal matrix H such that $P_0 = H^\top DH$ and $\|R_0 - R_0^{\text{hist}}\|_F \to$ min, where $R_0 = \Sigma_0^{-1} P_0 \Sigma_0^{-1}$ with $\Sigma_0 = \left(\text{diag}(P_0)\right)^{1/2}$ (see [3]). We report the so-found covariance matrix as

$$P_0 = \begin{pmatrix} 0.0233 & -0.0005 \\ -0.0005 & 0.0427 \end{pmatrix} . \tag{10}$$

Time-dependent, skew-symmetric matrices $Y_i(t)$ can be obtained by multiplying an arbitrary time-dependent function $g_i(t)$ with the generator G of $\mathfrak{so}(2)$, i.e. $Y_i(t) = g_i(t)G$ for $i = 0, 1, 2$. Experimenting with different functions we chose

$$g_0(t) = x_1 t \sin(x_2 t), \quad g_1(t) = x_3 + x_4 t, \quad g_2(t) = x_5 + x_6 t, \tag{11}$$

as they worked best regarding the given historical data. The parameters $x_1, \ldots, x_6 \in \mathbb{R}$ can be associated with possible degrees of freedom.

Structure-Preserving Euler-Maruyama Scheme
We solve (7) with the initial value and coefficients specified in the previous step by applying the following algorithm which is based on RKMK schemes for ODEs [5].

Algorithm *Divide the time interval* $[0, T]$ *uniformly into* J *subintervals* $[t_j, t_{j+1}]$, $j = 0, 1, \ldots, J-1$ *and define* $\Delta = t_{j+1} - t_j$ *and* $\Delta W_i \sim \mathcal{N}(0, \Delta)$. *Starting with* $t_0 = 0$ *and* $\Omega_0 = 0$ *these steps are repeated until* $t_{j+1} = T$*:*

1. *Let* P_j *be the approximation of* P_t *at time* $t = t_j$.
2. *Compute* Ω_1 *by applying the Euler-Maruyama scheme to the SDE* (6).
3. *Define a numerical solution of* (7) *as* $P_{j+1} = \exp(\Omega_1) P_j \exp(-\Omega_1)$. □

The computation of the correlation flow can be listed as an additional step:

4. Set $R_{j+1} = \Sigma_{j+1}^{-1} P_{j+1} \Sigma_{j+1}^{-1}$ with $\Sigma_{j+1} = \big(\operatorname{diag}(P_{j+1})\big)^{1/2}$.

Calibration
We calibrate the parameters $x_1, \ldots, x_6$ in (11) such that the mean squared error, $\frac{1}{N}\sum_{j=1}^{N}\big(f^{\text{hist}}(z_j) - f^{\text{flow}}(z_j)\big)^2$ is minimized, where $f^{\text{hist}}(z)$ and $f^{\text{flow}}(z)$ are the empirical density function of the historical data and the correlation flow, resp., estimated with the MATLAB function `ksdensity` at $N = 100$ equally spaced points.

Choosing $(x_1, x_2, x_3, x_4, x_5, x_6) = (6.22, -5.22, 9.88, -5.19, -0.62, -16.63)$ we computed a mean squared error of $9.57 \cdot 10^{-4}$. A corresponding plot that shows how well the density function of our correlation flow approximates the historical data can be found in Fig. 2.

4 Conclusion

We have presented an approach that shows that the correlation model of [6] can be extended such that the stochastic behaviour of correlations is included by modelling the isospectral flow as a SDE instead of an ODE. Moreover, we introduced a structure-preserving scheme that keeps the numerical solution of this stochastic isospectral flow on the correct manifold $\widehat{\mathrm{Sym}}(n)$. Lastly, we have seen that our methodology for the approximation of correlation matrices based on the stochastic isospectral flow works quite well. In future work one could extend our model such that more correlations ($n > 2$) are approximated. For this purpose, one could adjust the number of diffusion coefficients $Y_{i,t}$ and the time-dependent functions $g_i(t)$ or apply higher order methods.

References

1. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Springer (2006)
2. Malham, S.J.A., Wiese, A.: Stochastic Lie group Integrators. SIAM J. Sci. Comput. **30**(2), 597–617 (2008)
3. Muniz, M., Ehrhardt, M., Günther, M.: Approximating Correlation Matrices using Stochastic Lie Group Methods Mathematics **9**(1):94 (2021)
4. Muniz, M., Ehrhardt, M., Günther, M., Winkler, R.: Higher Strong Order Methods for Itô SDEs on Matrix Lie Groups. BIT Numerical Mathematics, 2022.
5. Munthe-Kaas, H.: High order Runge-Kutta Methods on Manifolds. Appl. Numer. Math. **29**, 115–127 (1999)
6. Teng, L., Wu, X., Günther, M., Ehrhardt, M.: A new methodology to create valid time-dependent correlation matrices via isospectral flows. ESAIM: Math. Model. Numer. Anal. **54**(2), 361–371 (2020)

Investigation of Darwin Model with Two Types of Coulomb Gauge Condition in Frequency-Domain Electromagnetic Finite-Element Method

Hiroyuki Kaimori, Takeshi Mifune, and Akihisa Kameari

Abstract In quasi-static electromagnetic field analysis, the Darwin model, which considers inductive and capacitive effects, has attracted much attention. Many previous methods require the additional scalar potentials for low-frequency (LF) stabilization, as low-frequency stabilization is essential to obtain a correct solution in broadband simulations. Additionally, it is necessary to solve the issues related to eddy current by considering the inductance and capacitance effects. In this paper, two new effective methods in the frequency domain are proposed for quasi-static electromagnetic finite-element method using an iterative matrix solver. The proposed methods are the Coulomb gauge condition applied without any additional scalar potential and the Coulomb gauge condition applied with redundant variable to improve numerical stability. The proposed methods can also be used for calculations using external electrical circuits. The numerical results verify the effectiveness of the two proposed methods.

1 Introduction

Although the quasi-static electromagnetic finite-element method (FEM) of the $\boldsymbol{A}$-ϕ formulation can solve issues related to eddy current, it cannot manage the capacitive effects because the formulation neglects displacement currents. Therefore, the Darwin model [1, 2] was investigated for the quasi-static $\boldsymbol{A}$-ϕ formulation [2, 5–9]. The Darwin model treats the dielectric as an electrostatic field, neglecting the displacement current. However, the electromagnetic field FEM of the $\boldsymbol{A}$-ϕ formulation of the Darwin model struggles to obtain a correct solution under

H. Kaimori (✉) · A. Kameari
Science Solutions International Laboratory, Inc., Tokyo, Japan
e-mail: kaimorih@ssil.co.jp; kamearia@ssil.co.jp

T. Mifune
Graduate School of Engineering, Kyoto University, Kyoto, Japan
e-mail: mifune.takeshi.3v@kyoto-u.ac.jp

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_60

ungauged conditions. This problem is known as low-frequency (LF) stabilization [3, 4]. Therefore, in the aforementioned methods, additional scalar variables are defined to improve the numerical stability. The LF-stabilization has been discussed numerically in terms of ill-conditioned coefficients, such as singular matrices. This may be why it is often solved using direct matrix solvers. However, the physical meaning of why the gauged condition requires an additional scalar variable has not been fully clarified.

In this paper, we clarify the reason for using the Coulomb gauge condition in the Darwin model of the $\boldsymbol{A}$-ϕ formulation. We propose two new methods for the Darwin model that are effective in the frequency domain: (1) by not defining additional variables and (2) by defining redundant variables. Note that the redundant variables are defined by satisfying the gauge condition, whereas in previous studies, additional variables are defined for numerical stability. Both methods satisfy the numerical stability requirement of LF-stabilization and provide the correct solution. The parallel connection model of an inductor and capacitor is analyzed and compared with the conventional eddy current analysis to verify the effectiveness of both methods.

2 Potential Formulation with Darwin Model

2.1 Darwin Model with Coulomb Gauge Condition

We consider the finite-element method of the $\boldsymbol{A}$-ϕ formulation using the Darwin model. According to Helmholtz's theorem, a vector field can be decomposed into two components: transverse and longitudinal. By applying this theorem to the electric field $\boldsymbol{E}$, the transverse component $\boldsymbol{E}_T$ is represented by the magnetic vector potential $\boldsymbol{A}$ as the induced electric field, and the longitudinal component $\boldsymbol{E}_L$ is represented by the electric scalar potential ϕ as the Coulomb electric field [1]. Therefore, $\boldsymbol{E}$ can be expressed in the frequency domain (by $j\omega$) as:

$$\boldsymbol{E} = \boldsymbol{E}_T + \boldsymbol{E}_L = -j\omega\boldsymbol{A} - \nabla\phi. \tag{1}$$

Applying Eq. (1) to the Darwin model, the constitutive equations: Ohm's law and the relation between $\boldsymbol{E}$ and electric flux density $\boldsymbol{D}$ are defined as follows:

$$\boldsymbol{J}_e = \sigma(\boldsymbol{E}_T + \boldsymbol{E}_L) = -\sigma(j\omega\boldsymbol{A} + \nabla\phi), \tag{2}$$

$$\boldsymbol{D} = \epsilon\boldsymbol{E}_L = -\epsilon\nabla\phi, \tag{3}$$

where σ is the conductivity and ϵ is the permittivity. Note that $\boldsymbol{D}$ is expressed in the relation $\boldsymbol{E}_L$. By substituting the constitutive equations into the Ampere-Maxwell equation and the continuity equation of the $\boldsymbol{A}$-ϕ formulation, we obtain the Darwin-Ampere-Maxwell equation and the Darwin continuity equation as follows:

$$\nabla \times (\nu \nabla \times \boldsymbol{A}) + \sigma(j\omega \boldsymbol{A} + \nabla\phi) + \epsilon(j\omega\nabla\phi) = \boldsymbol{J}_s, \tag{4}$$

$$-\nabla \cdot \{\sigma(j\omega \boldsymbol{A} + \nabla\phi) + \epsilon(j\omega\nabla\phi)\} = 0. \tag{5}$$

where ν and $\boldsymbol{J}_s$ denote the reluctivity and the source current, respectively. The displacement current term is approximated by $\partial_t \epsilon \boldsymbol{E}$, as the 2nd derivative of time is neglected, $\partial_t^2 \epsilon \boldsymbol{A} = 0$ [2, 5]. The practice of neglecting the $\epsilon\nabla\phi$ terms in Eqs. (4) and (5) is common in the $\boldsymbol{A}$-ϕ formulation of eddy current analysis, which can be solved precisely by iterative solvers without imposing the gauge condition. Although Eqs. (4) and (5) do not impose the gauge condition, the iterative solvers are generally significantly slow [5]. Unfortunately, we cannot obtain convergence solutions with IC-BiCGStabs solver. To avoid this, it may be necessary to apply artificial conductivity to the nonconductive region [2], which may lead to unexpected errors in the conductive region (the skin effect cannot be correctly described). To solve this problem, we consider imposing the Coulomb gauge condition [3, 7], which is formulated as follows:

$$\nabla \cdot \epsilon j\omega \boldsymbol{A} = 0. \tag{6}$$

Applying the Galerkin procedure to the weak forms in Eqs. (4)–(6) and imposing the Coulomb gauge condition on the Darwin model, the proposed $\boldsymbol{A}$-ϕ formulation is given by:

$$\int_{\Omega} \boldsymbol{N} \cdot \big\{\nabla \times \nu\nabla \times \boldsymbol{A} + \nabla \cdot \epsilon(j\omega\nabla\phi) - \boldsymbol{J}_s\} dV + \int_{\Omega_C} \boldsymbol{N} \cdot \{\sigma(j\omega\boldsymbol{A} + \nabla\phi)\big\} dV = 0, \tag{7}$$

$$-\int_{\Omega_C} W\{\nabla \cdot \sigma(j\omega\boldsymbol{A} + \nabla\phi)\} dV - \int_{\Omega} W\{\nabla \cdot \epsilon(j\omega\nabla\phi)\} dV = 0, \tag{8}$$

$$\int_{\Omega} W\{\nabla \cdot \epsilon(j\omega\boldsymbol{A})\} dV = 0, \tag{9}$$

where $\boldsymbol{N}$ and W are the edge and nodal test functions, and Ω and Ω_C denote the entire and conductive region, respectively. By constructing a matrix for the system of Eqs. (7)–(9) and symmetrizing the Darwin continuity equation by dividing by $j\omega$, the formulation can be obtained as:

$$\begin{pmatrix} K^{\nu}_{AA} + j\omega C^{\sigma}_{AA} & C^{\sigma}_{A\phi} + j\omega C^{\epsilon}_{A\phi} \\ C^{\sigma}_{\phi A} + j\omega C^{\epsilon}_{\phi A} & -\frac{j}{\omega} C^{\sigma}_{\phi\phi} + C^{\epsilon}_{\phi\phi} \end{pmatrix} \begin{pmatrix} A \\ \phi \end{pmatrix} = \begin{pmatrix} J_A \\ 0 \end{pmatrix}. \tag{10}$$

Here, A and ϕ are discrete vectors containing unknowns, and K^{ν}, C^{σ}, and C^{ϵ} denote discrete material matrices of reluctivity, conductivity, and permittivity, respectively. The subscripts AA, $A\phi$, and ϕA in the discrete material matrices indicate that they are associated with variables A and ϕ. Eq. (10) can provide the means to improve the convergence characteristics and generate the correct solutions. However, the number of iterations of the iterative solvers tends to increase for high frequency problems. To improve the convergence characteristics, additional scalar variables, such as electric scalar potential $\phi = \phi + \psi$ [4] are introduced, but their physical meaning is not clear. Therefore, we introduce the $\boldsymbol{A}$-ϕ formulation using redundant variables to improve the convergence characteristics of the iterative solvers. We define the redundant variable χ in the entire region using a similar general gauge function:

$$\boldsymbol{A} = \boldsymbol{A}' + (j\omega)^{-1}\nabla\chi, \quad \phi = \phi' - \chi. \tag{11}$$

Eq. (11) neglect the 2nd derivative of time for ϕ and $\boldsymbol{A}$. Substituting Eq. (11) into Eqs. (7)–(9), we obtain another $\boldsymbol{A}$-ϕ formulation, which uses redundant variables:

$$\int_{\Omega} \boldsymbol{N} \cdot \{\nabla \times \nu \nabla \times \boldsymbol{A}' + \nabla \cdot \epsilon(j\omega\nabla(\phi' - \chi)) - \boldsymbol{J}_s\} dV + \int_{\Omega_C} \boldsymbol{N} \cdot \{\sigma(j\omega\boldsymbol{A}' + \nabla\phi')\} dV = 0, \tag{12}$$

$$-\int_{\Omega_C} W\{\nabla \cdot \sigma(j\omega\boldsymbol{A}' + \nabla\phi')\} dV - \int_{\Omega} W\{\nabla \cdot \epsilon(j\omega\nabla(\phi' - \chi))\} dV = 0, \tag{13}$$

$$\int_{\Omega} W\{\nabla \cdot \epsilon(j\omega\boldsymbol{A}' + \chi)\} dV = 0. \tag{14}$$

By constructing a system matrix for Eqs. (12)–(14), symmetrizing the Darwin continuity equation by dividing by $j\omega$, and transposing the rows, the formulation is obtained as:

$$\begin{pmatrix} K_{AA} + j\omega C^{\sigma}_{AA} & C^{\sigma}_{A\phi} + j\omega C^{\epsilon}_{A\phi} & -j\omega C^{\epsilon}_{A\chi} \\ C^{\sigma}_{\phi A} + j\omega C^{\epsilon}_{\phi A} & -\frac{j}{\omega} C^{\sigma}_{\phi\phi} + C^{\epsilon}_{\phi\phi} & \\ -j\omega C^{\epsilon}_{\chi A} & & -C^{\epsilon}_{\chi\chi} \end{pmatrix} \begin{pmatrix} A' \\ \phi' \\ \chi \end{pmatrix} = \begin{pmatrix} J_A \\ 0 \\ 0 \end{pmatrix} \tag{15}$$

Note that the symmetric system matrix can be solved correctly by adding χ to Eq. (10). Noticeably, χ is not defined as the gauge condition or LF-stablization. Moreover, both formulations can be solved using iterative solvers by applying the Coulomb gauge condition.

2.2 Necessity of Coulomb Gauge Condition

The role of the Coulomb gauge condition in providing the correct solutions is discussed in this section. It is known that the $\boldsymbol{A}$-ϕ and $\boldsymbol{A}$ formulations of a full wave can be solved correctly. However, the $\boldsymbol{A}$-ϕ formulation of the Darwin model cannot neglect ϕ because the representation in Eq. (3) is defined for the Coulomb electric field. $\boldsymbol{D}$ of the $\boldsymbol{A}$-ϕ formulation for a full wave is defined as:

$$\boldsymbol{D} = \epsilon\boldsymbol{E} = -j\omega\epsilon\boldsymbol{A} - \epsilon\nabla\phi. \tag{16}$$

Additionally $\boldsymbol{D}$ in the Darwin Model is defined by Eq. (3). To express $\boldsymbol{D}$ in a manner that yields the correct solution in the Darwin model, the divergence in both Eqs. (3) and (16) must be equal:

$$\nabla \cdot j\omega\epsilon \boldsymbol{A} + \nabla \cdot \epsilon \nabla\phi = \nabla \cdot \epsilon \nabla\phi. \tag{17}$$

The first term on the left-hand side of Eq. (17) is the Coulomb gauge condition defined in Eq. (6). Therefore, the Coulomb gauge must be treated explicitly in the $\boldsymbol{A}$-ϕ formulation of the Darwin model.

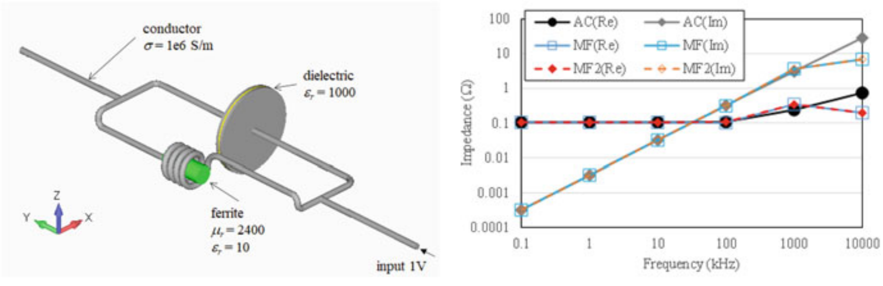
In addition, Gauss's law in the $\boldsymbol{A}$-ϕ formulation is represented in the quasi-electrostatic equation as:

$$\nabla^2 \cdot \epsilon \nabla\phi + \nabla \cdot j\omega\epsilon \boldsymbol{A} = \rho. \tag{18}$$

The 2nd term on the left-hand side of Eq. (18) is considered to be 0, which is exactly the Coulomb gauge condition in Eq. (6) for the electrostatic equation. Consequently, the electric field of the Darwin model solves the electrostatic field.

3 Numerical Results

A parallel connection model of an inductor and capacitor, in which the inductive effect is dominant at low frequencies and the capacitive effect appears at high frequencies, was analyzed. The model and parameters are shown in Fig. 1a. In this case, the capacitive effect appeared as an induced electric field in the dielectric region of the capacitor and ferrite. The boundary conditions were $\boldsymbol{B} \cdot \boldsymbol{n} = 0$, $\boldsymbol{E} \times \boldsymbol{n} = 0$ for both sides in the Y direction, and $\boldsymbol{H} \times \boldsymbol{n} = 0$, $\boldsymbol{D} \cdot \boldsymbol{n} = 0$, $\boldsymbol{J} \cdot \boldsymbol{n} = 0$ for the other boundary surfaces. An AC voltage of 1 V and a frequency of 100–10 MHz in 10-fold increments were applied to the terminals positioned in the



(a) (b)

Fig. 1 Parallel connection model of an inductor and capacitor. (**a**) Model and (**b**) impedance characteristics

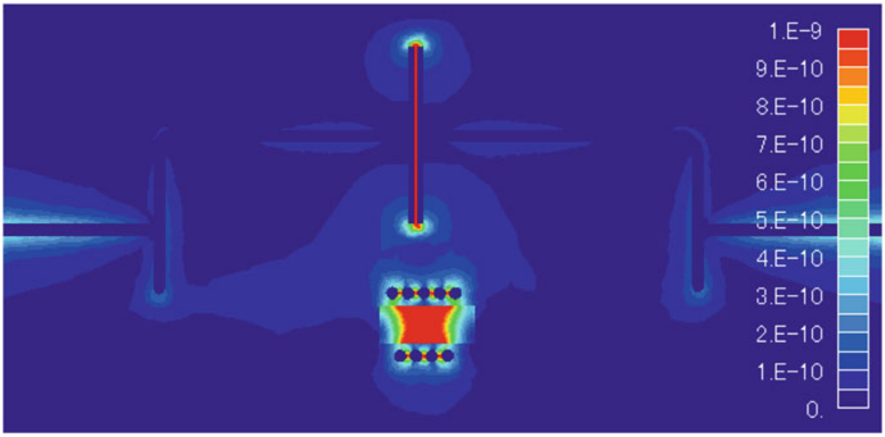


Fig. 2 The electric flux density distributions solved by the Darwin model without additional scalar variables at 10 MHz

Y-direction of the bottom surface by given ϕ. The output currents were calculated as the current passing through the bottom surface of the terminal. The IC-COCR solver was applied to solve the symmetrical matrix of equations. Figure 1b shows the impedance characteristics of the proposed Darwin model without additional scalar variables (MF), with the addition of redundant variables (MF2), and conventional eddy current analysis (AC), which includes the inductive effect only, for reference. Above 1 MHz, the results of the Darwin models MF and MF2 were the same, but different from AC because of phase shift caused by the capacitance effect. Figure 2 shows the electric flux density distributions at 10 MHz solved using the Darwin model without additional scalar variables. As can be seen, the electric flux density appears around the capacitor, ferrite, and between the coil turns. Below 100 kHz, the number of iterations for AC, MF, and MF2 was almost the same, averaging approximately 900. However, the MF increased to approximately 2000 iterations above 100 kHz. Therefore, the redundant variables may have improved the convergence characteristics. The reason for the increase in the number of iterations in MF will be discussed in future studies.

4 Conclusions

We presented the quasi-static electromagnetic $\boldsymbol{A}$-ϕ formulations of the Darwin model using two different approaches: (1) without additional variables and (2) with the addition of redundant variables. We explained the necessity of the Coulomb gauge condition for the Darwin model and showed that it is an electrostatic field approximation. It was verified that both method can calculate the electric field correctly and that a capacitive effect can be obtained using the parallel connection model of an inductor and capacitor.

References

1. Larsson, J.: Electromagnetics from a quasistatic perspective. Amer. J. Phys., 75, 3, 230–239 (2007)
2. Koch, S., Schneider, H. and Weiland, T.: A low-frequency approximation to the Maxwell equations simultaneously considering inductive and capacitive phenomena. IEEE Trans. Magn., 48, 2, 551–514 (2012)
3. Hiptmair, R., Kramer, F. and Ostrowski, J. M.: A Robust Maxwell Formulation for All Frequencies. IEEE Trans. Magn., 44, 6, 682–685 (2008)
4. Ostrowski, J. and Hiptmair, R.: Frequency-Stable Full Maxwell in Electro-Quasistatic Gauge. SIAM J. Sci. Comput., 43(4), B1008–1028. (2021)
5. Ho, S. L., Zhao, Y., Fu, W. N. and Zhou, P.: Application of Edge Elements to 3-D Electromagnetic Field Analysis Accounting for Both Inductive and Capacitive Effects. IEEE Trans. Magn., 52, 3, Mar. 7400504 (2016)
6. Zhao, Y. and Fu, W. N.: A Novel Coulomb-Gauged Magnetic Vector Potential Formulation for 3-D Eddy-Current Field Analysis Using Edge Elements. IEEE Trans. Magn., 53, 6, June, 9400704 (2017)
7. Zhao, Y. and Tang, Z.: A Novel Gauged Potential Formulation for 3-D Electromagnetic Field Analysis Including Both Inductive and Capacitive Effects. IEEE Trans. Magn., 55, 6, June, 7200905 (2019)
8. Jochum, M., Farle, O., and Dyczij-Edlinger, R.: A New Low-Frequency Stable Potential Formulation for the Finite-Element Simulation of Electromagnetic Fields. IEEE Trans. Magn., 51, 3, March, 7402304 (2015)
9. Badics, Z., Pavo, J., Bilicz, S. and Gyimothy, S.: Subdomain Perturbation Finite-Element Method for Quasi-static Darwin Approximation. IEEE Trans. Magn., 56, 1, June, 7503304 (2020)

Statistical and Machine Learning Methods for Automotive Spare Parts Demand Prediction

Tiago Carmo, Manuel Cruz, Jorge Santos, Sandra Ramos, Sofia Barroso, and Patrícia Araújo

Abstract Nors is a Portuguese group working on transport solutions. One of Nors companies is a wholesaler of automotive spare parts dealing with several hundreds of thousands of references, provided by different suppliers which have their own lead-time and order periodicity. Given the magnitude of the references set, the stock value and operational costs are non-negligible factors concerning their impact on the company operational results. Nors already has a mathematical prediction model for the spare parts ordering and management system. This work intends to improve the existing model through the application of Neural Networks, namely Long Short-Term Memory (LSTM) Neural Networks, both as a standalone prediction model and as a combination with the existent one. The results show that in fact there is an improvement with a consequent potential reduction in stock and warehouse costs.

1 Introduction

High number of references and suppliers cause a need for good prediction algorithms to manage stocks and orders and fulfil the maximum of the existing demand.

Nors group has a sales prediction algorithm which bases its prediction in 11 different methods [1]. To attempt to improve the existing algorithm, we implement recurrent neural networks [3], specifically, LSTMs [2]. LSTMs are widely used for time series prediction, reason for its application in this problem. The main difference between LSTMs and regular neural networks is the fact that LSTMs have the ability of capturing long-term temporal dependencies. Instead of a feed-forward approach,

T. Carmo · S. Barroso · P. Araújo
Nors Group, Porto, Portugal
e-mail: up201506221@edu.fc.up.pt; sbarroso@nors.com; paaraujo@nors.com

M. Cruz (✉) · J. Santos · S. Ramos
LEMA - Engineering Mathematics Laboratory, School of Engineering, Polytechnic of Porto, Porto, Portugal
e-mail: mbc@isep.ipp.pt; jms@isep.ipp.pt; sfr@isep.ipp.pt

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_61

in which the input is passed from one layer to the next one and so-on, recurrent neural networks feed the information to the different layers in loops. In a layer, each cell receives two inputs: the outputs of the previous layer and the vector of states from the current cell from the previous time step. The units of these recurrent neural networks also have a difference: they are gated recurrent units. That is, these gates can "decide" how much information is stored in each cell at each time step in order to keep the relevant information and avoid it to be "diluted" if we kept all the information from all inputs. In this work, LSTMs were used as regressors, predicting a single value from a sequence vector, in this case, one LSTM for each time series (each reference) in a total of 2565. They were used not only to predict the company sales, but also as predictors of the errors of the existing methods' forecasts. This last procedure constitutes the main novelty of the proposed work since, as far as we know, it has never been proposed before.

The preliminary parametrizations were implemented both in MATLAB and Python, to compare these languages in terms of accuracy and running times. The final experiments were performed only in MATLAB and the simulation results show that this approach is valid and that it improves the existing prediction algorithm.

Following this introduction, in the next section we describe the data and the methodology, in Sect. 3 we present the computational results on a subset of Nors real data and in the final section we discuss the results and draw some conclusions.

2 Materials and Methods

The dataset used in the simulations is a small subset from one supplier of the company. It has 2565 reference entries and a 50-month time history. These references represent different types of automotive spare parts from a single supplier and with a broad range of frequency sales. The dataset also includes the original forecasts (corresponding to the output of the 11 methods) made by the Nors sales prediction algorithm. The dataset matrix is very sparse, containing many references with few or even no sales. The heterogeneity of the dataset is also reflected by the fact that there are references that have a lot of sales but without following any apparent pattern, and a lot of low-rotation ones. These factors make it difficult to choose the global parameters of the LSTMs, because, despite being individually trained, they all share the same configuration (same parameters).

The data were split addressing 47 months for the training set and 3 months for testing. After some preliminary tests, we obtained the final architecture and hyperparameters for the LSTMs. It was used a single hidden layer, the Adam [4] optimizer (an extension to the stochastic gradient descent method) and a piecewise learn rate schedule with a drop factor of 0.2 and a drop period of 25 epochs. The search space for the number of hidden units was $60 + 10k,\ k = 0, 1, \ldots, 6$, for the learning rate $0.0005+0.0001k, k = 0, 1, 2, 3, 4, 0.001+0.001k, k = 0, 1, \ldots, 9$ and $0.01+0.01k, k = 0, 1, 2, 3, 4$, and for the number of epochs $60+20k, k = 0, 1, 2, 3$.

After the tests, the chosen values for the hyperparameters were 60 hidden units, 0.005 for the learning rate and 100 epochs. The values chosen for these hyperparameters were the ones that resulted in much lower running times, with good performance. This performance was measured as the percentage of references that had a lower error than the original Nors prediction.

These parametrizations were implemented both in MATLAB and Python for performance testing (running times and accuracy). Since there was no significant differences between both methods and since the existing methods are already implemented in MATLAB, it was decided to use this platform for the final simulations presented in the next section. Prophet [5] was also used to serve as a comparison [6].

3 Results

3.1 *LSTMs as Sales Predictors and as Predictors of the 11 Nors Original Methods' Errors*

To perform sales predictions, LSTMs were trained, and each month took about 45 min. of running time to get the results (around 1 sec. per reference). We compared the LSTM results with the method chosen by the Nors algorithm, both with exact results and with them rounded to the nearest integer. When considering the exact results, 33.1 % of the references have a better result than the method chosen by the Nors algorithm as the best one and 51.5 % of the references have a worse result. When considering the results rounded to the nearest integer, 11.4 % of the references are better than the best original one and 18.7 % are worse (Table 1, first column).

To predict the errors of the 11 Nors original methods, the training of the LSTMs took about 21 hours, which gives about 42 min. for each month and each method (around 1 sec. per reference). The errors are calculated by $Err_i = \hat{S}_i - S$, $i = 2, 3, \ldots, 11$, where Err_i represents the error of method i's prediction, $\hat{S}_i$ the method i's monthly sales prediction and S the real monthly sales for each given reference.

After supplying the LSTMs with these errors and obtaining their predictions, the original methods are adjusted in order to incorporate a prediction of its own errors. That is done according to $\hat{S}_i* = \hat{S}_i - \widehat{Err}_i$, $i = 2, 3, \ldots, 11$, where $\hat{S}_i*$ represents the new method i's prediction, now with a neural network prediction of its original error, and $\widehat{Err}_i$ represents the prediction of the error of method i's prediction. The corresponding results are presented in Table 1.

Table 1 Percentage of cases where each LSTM method obtained better (B) or worse (W) results than the method chosen by the algorithm implemented in Nors Group [1]

Method		NN	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
Exact	B	33.1	12.1	85.0	21.3	27.3	24.4	16.6	34.9	26.9	21.8	26.7
	W	51.5	19.8	8.3	24.8	27.1	25.5	29.2	28.3	25.0	33.2	25.0
Rounded	B	11.4	7.6	80.7	7.5	6.4	7.1	8.2	9.0	8.0	9.0	4.0
	W	18.7	4.4	3.2	2.9	5.6	5.2	5.0	7.6	5.9	7.2	4.5

3.2 Accuracy Metrics

To better evaluate the performance of the LSTMs according to the business perspective, some accuracy metrics were used. To do so, the methods involved were first partitioned in the following subsets:

- C1: Set of the 11 methods already implemented in the Nors algorithm.
- C2: LSTMs as sales predictor.
- C3: Set C1 plus set C2, consisting in a total of 12 methods.
- C4: Set of all the 22 methods involved, the 11 in C1, C2 and the 10 LSTMs applied to the errors of the C1 methods (except for method 1).
- C5: Prophet [5]: a forecasting procedure for time series developed by Facebook.

Prophet is designed to be fully automatic, although the user may tune several parameters. In this work we made several experiments, using different values for the parameters seasonality_mode (fits additive or multiplicative seasonality) and interval_width (confidence interval parametrization). Concerning the first parameter, the overall performance was similar in both options, except for the computation time where the multiplicative model took around 5 times more than the additive version.

In average, each reference of each original method is computed in 1.15×10^{-6} seconds, each reference of each LSTMs' method takes 1.00 seconds to compute and Prophet (additive model) takes about 1.38 seconds to compute each reference's prediction. The accuracy metrics considered were the following:

- Hits: Percentage of references where prediction matched the exact monthly sale.
- Service Level: Percentage of references where the prediction plus a safety stock was enough to satisfy the demand.
- Stock Out: Percentage of references where the prediction plus a safety stock was not enough to satisfy the demand.
- Excess: Percentage of references overordered.
- Stock Out Parts: Percentage of parts that did not satisfy the demand.
- Excess Parts: Percentage of parts that were ordered in excess.

Nors Group also has implemented a safety stock [1] which is a dynamic approach that evaluates, for a given reference in a certain month, the minimum stock needed in order to satisfy a certain service level. This evaluation takes into consideration the medium error of the reference predictions from the last 12 months and a management decision parameter (α) that is settled using the errors of the previous forecasts, as defined in [1, equation (20)]. This parameter was set with a value $\alpha = 0.04$ that would result in a service level of around 96%.

All the references' demand was forecasted for the following months: November 2019; January, October, November and December 2020; January and February 2021, and the corresponding accuracy metrics were computed (see Table 2).

The accuracy metric of forecasting the monthly sales' exact value (Hits), ranks all the bundles of methods (C1 to C4) as outperforming the Prophet algorithm (C5), with C2 having a slightly worse performance than C1, C3 and C4.

Table 2 Percentages of the different accuracy metrics for each month and each set of methods

	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Subsets	Hits					Service Level				
Nov. 2019	59.3	56.0	59.1	59.1	49.0	95.6	96.3	95.7	95.7	85.0
Jan. 2020	58.2	54.6	58.0	58.3	50.0	96.5	97.1	96.5	96.4	88.0
Oct. 2020	61.3	58.8	61.4	60.9	52.0	97.0	97.2	96.9	96.9	92.0
Nov. 2020	64.0	61.4	63.7	62.7	53.0	97.5	97.6	97.5	97.5	93.0
Dec. 2020	64.3	61.6	64.1	63.5	52.0	97.5	97.7	97.5	97.3	94.0
Jan. 2021	71.2	70.3	71.4	71.1	57.0	99.5	99.6	99.6	99.5	95.0
Feb. 2021	67.5	61.8	67.1	65.9	59.0	97.5	97.8	97.7	97.5	95.0
Average	63.69	60.64	63.54	63.07	53.14	97.30	97.61	97.34	97.26	91.71
	Stock out					**Excess**				
Nov. 2019	4.4	3.7	4.3	4.3		81.4	81.9	81.5	81.3	
Jan. 2020	3.5	2.9	3.5	3.6		82.7	84.0	82.8	82.6	
Oct. 2020	3.0	2.9	3.1	3.1		81.2	81.3	81.1	80.8	
Nov. 2020	2.5	2.4	2.5	2.5		82.1	82.3	82.1	82.1	
Dec. 2020	2.5	2.3	2.5	2.7		81.3	81.8	81.3	81.1	
Jan. 2021	0.5	0.4	0.4	0.5		83.4	82.9	83.4	82.7	
Feb. 2021	2.5	2.2	2.3	2.5		79.3	79.3	79.3	78.7	
Average	2.70	2.40	2.66	2.74		81.63	81.93	81.64	81.33	
	Stock out parts					**Excess parts**				
Nov. 2019	7.7	9.4	7.6	8.2		148.1	162.8	147.8	147.0	
Jan. 2020	4.7	3.7	5.1	5.8		173.6	180.4	172.6	174.0	
Oct. 2020	4.3	3.7	4.3	4.2		192.9	220.2	194.1	195.5	
Nov. 2020	5.3	4.5	5.3	5.3		208.7	243.0	213.4	214.2	
Dec. 2020	6.2	3.7	6.3	6.7		174.3	199.2	175.2	174.7	
Jan. 2021	26.2	24.4	26.1	26.2		130.1	143.7	131.6	132.2	
Feb. 2021	4.2	4.8	3.9	4.7		202.0	244.9	223.2	224.1	
Average	8.37	7.74	8.37	8.73		175.67	199.17	179.70	180.24	

Considering the Service Level, it can easily be seen that (apart from Prophet, C5) all the sets fulfil the 96% service level intended by Nors Group. It seems important to notice that, even when adjusting the Prophet confidence interval parametrization to 99%, the results return a Service Level of 93.4%, which is significantly lower than the Nors threshold of 96%. As such, C5 was not included in the remaining tables.

Looking at the Stock Out we can see that the LSTMs as sales predictors (C2) as well as when adding them to C1 (C3) outperform C1. Regarding the Excess Stock, a clear improvement is noted with all the methods together (C4), having the lowest value of references overordered. However, it is important to note that the quantity ordered to the supplier is a function of the predicted value. In fact, the orders are dependent of several factors as it may be seen in [1]. Moving on to the percentage of Stock Out Parts there is a clear similarity between C1 and C3 (having the same average percentage), but C2 is performing clearly better than the other bundles.

Finally, when looking at the Excess Parts, adding LSTMs' methods does not seem to improve so much to the Nors implemented algorithm.

4 Discussion and Conclusions

Setting the management decision parameter $\alpha = 0.04$ as described in Sect. 3.2 results in a Service Level very close to what was intended (96%) for all sets of methods except for Prophet (C5), reason why its results were not presented for most of the accuracy metrics.

When considering the single method C2 (LSTM) and comparing with C1 which contains the 11 different existent methods, they have similar performances (sometimes even outperforming C1), that is, the networks as sale predictors performs similarly to the best one out of 11 methods, which is a very good indicator of the strong performance of the LSTMs in this context. Also, adding C2 to C1 (when we use LSTMs to correct the existent methods prediction), set C3, the percentage of exact matches (Hits) are similar to C1 and the percentage of references and pieces in excess are reduced, and also slightly increasing the Service Level. Finally, all of the deterministic and LSTMs methods (set C4) also reduces the Excess stock as well as the service level, while still reaching the 96 % Nors mark for this supplier.

All of these results point to an improvement to the current algorithm, potentially saving the company in stock and warehouse expenses. As future work, we will try to create subgroups of references to use specialized LSTMs for each of them hoping to further improve the obtained results. We are also confident on the possibility of using this approach in other similar problems from industry.

References

1. Cruz, M.B., Ramos, S.F., Pina, M. & Costa, R. (2021). Order and Stock Costs Optimization in an Automotive Spare Parts Wholesaler. *Progress in Industrial Mathematics: Success Stories: The Industry and the Academia Points of View*, 5, 145
2. Hochreiter, S., Schmidhuber, J. Long short-term memory. Neural Computation 9(8) (1997), 1735–1780.
3. Haykin, S. Neural networks and machine learning. Pearson Education (2009) 147139-9.
4. Diederik, K., Ba, J. (2014). Adam: A Method for Stochastic Optimization, ICLR2014.
5. Prophet, https://facebook.github.io/prophet/. (Accessed on 20/01/2022).
6. Shah V. A Comparative Study of Univariate Time-series Methods for Sales Forecasting, University of Waterloo, 2019, Canada.

The Parareal Algorithm and the Sparse Grid Combination Technique in the Application of the Heston Model

Anna Clevenhaus, Matthias Ehrhardt, and Michael Günther

Abstract The sparse grid combination technique is an efficient method to reduce the curse of dimensionality for high-dimensional problems, since it uses only selected grids for spatial discretization. To further reduce the computational complexity in the temporal dimension, we choose the Parareal algorithm, a parallel-in-time algorithm. For the coarse and fine solvers in time, we use an efficient implementation of the Alternating Direction Implicit (ADI) method, which is an unusual choice due to the larger computational cost compared to the usual choice of one-step or Runge-Kutta methods. In this paper we combine both approaches and therefore obtain a even more efficient computational method for parallelism. The application problem is to determine a fair price of a Put option using the Heston model with correlation. We analyze this model as an example to illustrate this advantageous combination of the sparse grid with the Parareal algorithm. Finally, we present further ideas to improve this advantageous combination of methods.

1 American Option Pricing Under the Heston Model

The payoff function for a Put option with a predefined strike K and the price for the underlying asset S is given by

$$\phi(S) = \max(K - S, 0).$$

The Heston model [4] describes the dynamics of the asset price S and the variance ν which is by definition the square of the volatility of the asset price. To price an American put option using the Heston model, we have to solve a free boundary value problem. We seek for $\big(P(S,\nu,t), S_f(t), \nu_f(t)\big)$ in $t \in [0, T]$, where $S_f(t)$ and $\nu_f(t)$ are the free boundary values at time t and $P(S,\nu,t)$ fulfills

A. Clevenhaus (✉) · M. Ehrhardt · M. Günther
Bergische Universität Wuppertal, Wuppertal, Germany
e-mail: clevenhaus@uni-wuppertal.de; ehrhardt@uni-wuppertal.de; guenther@uni-wuppertal.de

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_62

$$P(S,\nu,t)=\phi(S)\quad\text{for}\quad S\le S_f(t),\qquad P(S,\nu,t)>\phi(S)\quad\text{for}\quad S>S_f(t).$$

The differential operator for $P(S,\nu,t)$ is given by

$$\mathcal{L}[P]=\frac{1}{2}\nu S^2\frac{\partial^2 P}{\partial S^2}+\rho_{S\nu}\sigma_\nu S\nu\frac{\partial^2 P}{\partial S\partial\nu}+\frac{1}{2}\sigma_\nu^2\nu\frac{\partial^2 P}{\partial\nu^2}+rS\frac{\partial P}{\partial S}+\kappa_\nu(\nu-\mu_\nu)\frac{\partial P}{\partial\nu}-rP,$$

where r is the interest rate, κ_ν is the mean-reversion rate, σ_ν is the volatility-of-variance and μ_ν is the long-term mean of the variance ν. As for the variance $\nu>0$ holds, the Feller condition $2\kappa_\nu\mu_\nu>\sigma_\nu$ has to be fulfilled. The correlation between S and ν is denoted by $\rho_{S\nu}\in[-1,1]$. After time reversal $\tau=T-t$, the differential operator has to fulfill the inequality

$$\frac{\partial P}{\partial\tau}-\mathcal{L}[P]\ge 0$$

and the initial condition

$$P(S,\nu,0)=\phi(S),\qquad S>S_f(0).$$

At the boundaries of the asset, the payoff function, and at the boundaries of the variance, the inequality of the differential operator has to be fulfilled. To avoid an explicit computation of the free boundary value problem, we apply an operator splitting and recast the problem into a *linear complementarity problem* with an auxiliary variable λ [7]

$$\begin{cases}\frac{\partial P}{\partial\tau}-\mathcal{L}[P]=\lambda,\\ \lambda\ge 0,\ P-\phi(S)\ge 0,\quad \big(P-\phi(S)\big)\lambda=0.\end{cases}\tag{LCPλ}$$

In this mixed formulation of the LCP, λ plays the role of a Lagrange multiplier.

2 The Sparse Grid Combination Technique

The sparse grid idea is motivated to reduce the curse of dimensionality for solving PDEs [1]. Let $\mathbf{x}\in\Omega_2=[0,1]^2$ be defined by the multi-indices

$$\mathbf{l}=(l_1,l_2)\in\mathbb{N}_0^2,\quad \mathbf{j}=(j_1,j_2)\in\mathbb{N}_0^2,\quad \mathbf{N}=(N_1,N_2)=(2^{l_1},2^{l_2}).\tag{1}$$

such that we can define a tensor based grid $\Omega_\mathbf{l}$ whose grid nodes are given by

$$\mathbf{x}_{\mathbf{l},\mathbf{j}}=(x_{l_1,j_1},x_{l_2,j_2})\quad\text{for}\quad j_1=0,\dots,\mathrm{N}_1\quad\text{and}\quad j_2=0,\dots,\mathrm{N}_2.$$

The mesh width defined by this grid is $h = (2^{-l_1}, 2^{-l_2})$. If for some applications being sensitive to disordered grids, the difference between l_1 and l_2 is to high, we obtain these kind of grids and therefore avoidable modelling errors. To avoid those errors, we set a minimum for $l_i > l_{\min} = 3$, s.t. each spacial direction has at least 9 grid points. Let u be the continuous solution on Ω_2 and $u_{\mathbf{l}}$ the discrete solution on $\Omega_{\mathbf{l}}$ with $l = (l_1, l_2)$. The hierarchical surplus of $u_{\mathbf{l}}$ is denoted by

$$\delta(u_{\mathbf{l}}) = u_{\mathbf{l}} - u_{\mathbf{l}-e_1} - u_{\mathbf{l}-e_1} + u_{\mathbf{l}-e_1-e_2} \text{ with } e_1 = (1,0)^\top,\ e_2 = (0,1)^\top,$$

where w_1 only depends on h_1, w_2 only on h_2 and h_1 and h_2 are independent from each other. Further w_1, w_2, $w_{1,2}$ are bounded. Based on the error splitting

$$u - u_{\mathbf{l}} = h_1^2 w_1(h_1) + h_2^2 w_2(h_2) + h_1^2 h_2^2 w_{1,2}(h_1, h_2),$$

we derive the error spitting of the hierarchical surplus

$$\delta(u - u_{\mathbf{l}}) = \mathcal{O}\left(h_1^2 h_2^2\right) = \mathcal{O}\left(2^{-2|\mathbf{l}|_1}\right).$$

For the highest information gain for the sparse grid solution u_n^s of level $n = |\mathbf{l}|_1$, the sparse grid combination technique

$$u_n^s = \sum_{|\mathbf{l}|_1 \le n} \delta\left(u_{\mathbf{l}}\right) = \sum_{|\mathbf{l}|_1 = n} u_{\mathbf{l}} - \sum_{|\mathbf{l}|_1 = n-1} u_{\mathbf{l}}$$

is derived by the combination of the hierarchical surplus and the error splitting. Since the sparse grid combination technique is developed on Ω_2, we define $x = (y, z) \in [0,1]^2$ and obtain $S \in [S_{\min}, S_{\max}]$ and $\nu \in [\nu_{\min}, \nu_{\max}]$ by using the following transformation

$$\psi^{-1}(y) = S_0 + \alpha \cdot \sinh(y \cdot (c_2 - c_1) + c_1),$$
$$c_1 = \sinh^{-1}\left(\frac{S_{\min} - S_0}{\alpha}\right), \qquad c_2 = \sinh^{-1}\left(\frac{S_{\max} - S_0}{\alpha}\right),$$

where α describes the degree of the non-uniformity of the grid and $P(S_0, \nu_0, T)$ denotes the option price which we are interested in. If α is small, we obtain a highly non-uniform grid and else wise the non-uniformity aspires to a uniform grid. For z we use the transformation analogously. Using finite difference stencils of second order, the semi-discrete *partial differential complementarity problem (PDCPλ)*

$$\frac{\partial P}{\partial \tau} = FP(\tau) + \lambda(\tau), \quad P(\tau) \ge \phi\big(\psi^{-1}(y)\big), \quad \Big(P(\tau) - \phi\big(\psi^{-1}(y)\big)\Big)^\top \lambda(\tau) = 0,$$

is derived.

3 Temporal Discretization and the Parareal Algorithm

We discretize the time uniformly, using $\Delta_\tau = T/N_t$ we obtain the temporal time points $\tau_k = k \cdot \Delta_\tau$ with $k = 0, \dots, N_t$. With u^k describing the discrete solution at time step τ_k and g describing the discrete payoff value, we gain the fully *discrete linear complementarity problem*, cf. [7]

$$\mathcal{T}(u^k, \lambda^k, \tau^k) = \begin{cases} u^{k+1} = Au^k + \Delta_\tau \lambda^k, \\ \lambda^{k+1} \geq 0, u^{k+1} \geq g, (\lambda^{k+1})^\top (u^{k+1} - g). \end{cases} \tag{DLCPλ}$$

Within this problem, we have to solve two separate problems. In the first step a system of linear equations has to be solved and in the second one a variable update is done. The system of equations is solved by the modified Craig-Sneyd scheme with the additional parameter λ

$$\begin{cases} Y_0 = u^k + \Delta_\tau \mathcal{A}(\tau^k, u^k) \boxed{+\Delta_\tau \lambda^k}, \\ Y_i = Y_{i-1} + \theta \Delta_\tau \left(\mathcal{A}_i(\tau^k, Y_i) - \mathcal{A}_i(\tau^k, u^k)\right), \quad i = 1, 2, \\ \hat{Y}_0 = Y_0 + \theta \Delta_\tau \left(\mathcal{A}_0(\tau^k, Y_0) - \mathcal{A}_0(\tau^k, u^k)\right) \\ \tilde{Y}_0 = \hat{Y}_0 + (\frac{1}{2} - \theta) \Delta_\tau \left(\mathcal{A}(\tau^k, \hat{Y}_0) - \mathcal{A}(\tau^k, u^k)\right) \\ \tilde{Y}_i = \tilde{Y}_{i-1} + \theta \Delta_\tau \left(\mathcal{A}_i(\tau^k, Y_i) - \mathcal{A}_i(\tau^k, u^k)\right), \quad i = 1, 2, \\ \tilde{u}^{k+1} = \tilde{Y}_2, \end{cases}$$

where $\mathcal{A}_0$ is the operator for the mixed derivatives, $\mathcal{A}_1$ the operator of the derivatives of the first coordinate direction, $\mathcal{A}_2$ the operator of the derivative of the second direction and $\mathcal{A}$ the sum of all operators. The operators are defined due to the underlying pricing model, the Heston model. An improved way of implementation of the ADI schemes is used [9], where the computation is based on matrices instead of vectors and thus reduces the computational effort as redundant computations are avoided. Since numerical results show $N_1 - 2N_2 = 2^{l_1} - 2 \cdot 2^{l_2} = 0$ is a feasible choice [5], we apply additional restrictions to $\mathbf{l}$ [2]. The restrictions can vary from the strict condition $l_1 > l_2$ being fulfilled for every single sparse grid to a softer condition where $\max l_1 > \max l_2$ holds. We focus on the strict difference between l_1 and l_2 and introduce the parameter $\mathbf{l}_{\text{diff}} = \min(l_1 - l_2)$. The second step, the variable update can be done component wise by applying

$$\begin{cases} u^{k+1} &= \max(\tilde{u}^{k+1} - \Delta_\tau \lambda^k, u^0), \\ \lambda^{k+1} &= \max(0, \lambda^k + (u^0 - \tilde{u}^{k+1})/\Delta_\tau) \end{cases}.$$

As $u^0 = \psi(\phi(S))$, we set λ^0 as the zero vector.

The Parareal algorithm is an iterative parallel-in-time method and can be viewed as either a multigrid method or a multiple shooting method [8]. Within the iterative procedure, two different solvers are used. For both solvers, we consider the temporal operators as previously described. Both solvers themselves converge to the exact solution. The difference between the fine and coarse solvers is based on the considered spatial grid, the fine solver $\mathcal{F}$ solves the problem on u_n^s with $N_{\mathcal{F}}$ time steps and the coarse solver $\mathcal{G}$ on u_{n-1}^s with $N_{\mathcal{G}}$ time steps. We initialize the algorithm by introducing N_τ equal time slices, s.t. $\tilde{\tau}_p = [(p-1)\cdot\Delta_{\tilde{\tau}}, p\cdot\Delta_{\tilde{\tau}}]$, where $\Delta_{\tilde{\tau}} = \frac{T}{N_\tau}$. The initial value for the first time slice is always given by the initial condition. The initial guess for each time slice is calculated by the coarse solver. Since the fast solver solves one time slice in each iteration, after at least N_τ iterations the exact solution computed by $\mathcal{F}$ would be obtained. Therefore the maximum number of iterations J must be much smaller than N_τ. After initialization, the iterative procedure begins. First, the fine solver computes in parallel the solution of each time slice with the initial values. Let u_i^j be the discrete solution to the time slice $\tilde{\tau}_i$ at the j-th iteration. A serial correction step over all time slices follows

$$u_{i+1}^{j+1} = \mathcal{G}(u_i^{j+1}, \tilde{\tau}_i, \tilde{\tau}_{i+1}) + \mathcal{F}(u_i^j, \tilde{\tau}_i, \tilde{\tau}_{i+1}) - \mathcal{G}(u_i^j, \tilde{\tau}_i, \tilde{\tau}_{i+1}).$$

4 Numerical Results

In this section, we analyze the effect of reducing the grid resolution in the volatility direction on the accuracy as well as the application of the Parareal algorithm to the run time. We consider the following set of parameters

$$T = 0.25,\ K = 10,\ \rho_{S\nu} = 0.1,\ r = 0.1, \kappa_\nu = 5,\ \mu_\nu = 0.16,\ \sigma_\nu = 0.9,\ J = 3, l_{\min} = 3,$$

$$|\mathbf{l}|_1 = 12,\ S \in [0, 3K],\ \nu \in [0, 3], \alpha_S = \alpha_\nu = 2, N_\tau = 16,\ N_{\mathcal{F}} = 100,\ N_{\mathcal{G}} = 25.$$

This financial parameter set is often used and therefore is chosen to gain a comparison for results [3]. Table 1 contains the computed Put option prices for different grid resolutions, for each resolution the results are very close to the reference values obtained in [3]. Further we get to know that even for very small volatility values and a high reduction in resolution the results are comparable to the sparse grid solution containing also solutions with $l_1 = l_2$, which requires almost twice the amount of grid points as the restricted sparse grids and thus twice the computational time.

Figure 1 shows the run time results for different parallel processors using the same parameter set as before, but using different $|\mathbf{l}|_1$ values. We observe that the sparse grid technique is more efficient than the combination with the Parareal algorithm, due to increased communication time. To underline this fact, we observe that the runtime increases almost linearly with the number of processors. Note, that we choose $\mathbf{l}_{\text{diff}} = 0$ for a fair comparison, as the increase of $\mathbf{l}_{\text{diff}} > 1$ is only suitable for the parareal algorithm. Using such a increased sparse grid as underling grid

Table 1 Solution values for the different spot asset prices for the parameter sets compared to reference values computed by the Parareal algorithm using sparse grids

		$\nu_0 = 0.0625$					
S_0		8	9	10	11	12	
[3]		2.0000	1.1081	0.5204	0.2143	0.0827	Grids
$\mathbf{l}_{\text{diff}}$	0	2.0000	1.1078	0.5202	0.2138	0.0821	13
	1	2.0000	1.1078	0.5202	0.2138	0.0821	11
	2	2.0000	1.1075	0.5202	0.2138	0.0821	9
	3	2.0000	1.1076	0.5201	0.2137	0.0821	7

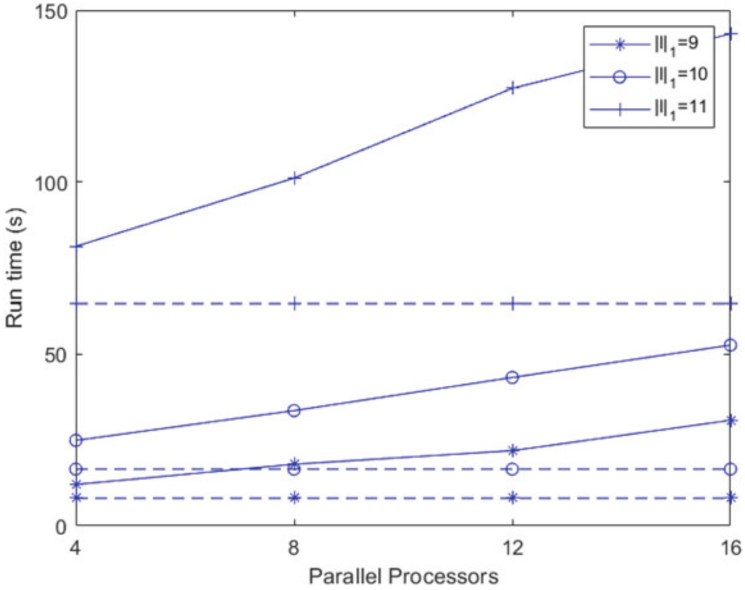


Fig. 1 The dashed line corresponds to the constant serial run time using sparse grids and the solid line represents the run time for the Parareal Algorithm with sparse grids with 4, 8, 12 and 16 parallel processors and $\mathbf{l}_{\text{diff}} = 0$

structure for the parareal algorithm, we would obtain a smaller runtime. This is only one of further improvement strategies which have to be applied to obtain a benefit even for smaller problems.

5 Conclusion and Outlook

The numerical results show that even the additional restriction $l_1 > l_2$ with $\mathbf{l}_{\text{diff}}$ large, which leads to a high resolution reduction in the volatility direction is feasible. To obtain better results for using the Parareal algorithm in combination with the sparse grid approach, we need to further improve the resulting algorithm. Fortunately, beneath the idea of using $\mathbf{l}_{\text{diff}} > 1$ there are two ways to reduce the computational cost. The first idea is based on the structure of the sparse grid combination technique. Since in the presented approach all sparse grids of level $|\mathbf{l}|_1 - 1$ have to be computed by the fine and the coarse solver, we can easily reduce the overhead by reusing the results. The second is based on parallelizing the computation of the sparse grids within the coarse solver, since they can each be computed independently.

References

1. H. Bungartz and M. Griebel, Sparse Grids, Cambridge University Press, 2004, 1–123.
2. A. Clevenhaus, M. Ehrhardt, and M. Günther, An ADI Sparse Grid method for pricing efficiently American Options under the Heston model, to appear: Adv. App. Math. Mech., (2021).
3. T. Haentjens and K. J. in't Hout, ADI schemes for pricing American Options under the Heston model, Appl. Math. Fin., 22 (2013), 207–237.
4. S.L. Heston, A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options, In: Review of Financial Studies 6 (2) (1993), 327–343.
5. K.J. in't Hout and S. Foulon, ADI Finite Difference Schemes for Option Pricing in the Heston Model with Correlation, Int. J. Numer. Anal. Mod., 7 (2010), 303–320.
6. K.J. in't Hout and B. Welfert, Unconditional stability of second-order ADI schemes applied to multi-dimensional diffusion equations with mixed derivative terms, Appl. Numer. Math., 59(3-4) (2009), 677–692.
7. S. Ikonen and J. Toivanen, Operator splitting methods for pricing American options under stochastic volatility, Numer. Math., 113(2) (2009), 299–324.
8. J.-L. Lions, Y. Maday, and G. Turinici, Résolution d'EDP par un schéma en temps "pararéel", C.R.A.S. Sér. I Math., 332(7) (2000), 661–668.
9. L. Teng and A. Clevenhaus. Accelerated implementation of the ADI schemes for the Heston model with stochastic correlation. J. Comput. Sci., 36 (2019), 101022.
10. C. Zenger, Sparse Grids, Technical Report, Institut für Informatik, Technische Universit"at München, October 1990.

A Higher-Order NSFD Method for a Simple Growth Model in the Chemostat

Fawaz K. Alalhareth and Hristo V. Kojouharov

Abstract Accurate numerical methods that also preserve the important properties of dynamical systems are essential, especially when approximating systems in science and engineering. In this paper, we analyze a simple growth model in the chemostat and present a new higher-order nonstandard finite difference (NSFD) method for it, which is positivity-preserving, elementary stable, and also of second-order accuracy. A set of numerical simulations is also presented to support the theoretical results.

1 Introduction

Nonstandard finite difference (NSFD) methods are widely used to solve many problems in science and engineering. The NSFD methods approximate the solutions of continuous problems and preserve some critical properties of the exact solutions, such as positivity and the local stability of the equilibria, among others. A methodology for designing positive and elementary stable nonstandard (PESN) numerical methods [2] was recently proposed for solving general autonomous systems with positive solutions [7]; however, they are only of first-order accuracy. More recently, a methodology for constructing second-order NSFD methods has been developed for one-dimensional differential equations [3, 4]; however they are only elementary stable [1] and do not preserve the positivity of solutions. This work

F. K. Alalhareth (✉)
The University of Texas at Arlington, Department of Mathematics, The University of Texas at Arlington, Arlington, TX, USA

Najran University, Department of Mathematics, Najran University, Najran, Saudi Arabia
e-mail: fawaz.alalhareth@mavs.uta.edu

H. V. Kojouharov
The University of Texas at Arlington, Department of Mathematics, The University of Texas at Arlington, Arlington, TX, USA
e-mail: hristo@uta.edu

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_63

proposes and analyzes a new second-order NSFD method for a simple growth model in the chemostat. T he NSFD method can be also applied to other two-dimensional autonomous differential equations in science and engineering.

The paper is organized as follows. In Sect. 2, we analyze a modification of the simple chemostat model, which incorporates the constant input and death of bacteria. Next, we present a new positivity-preserving, elementary stable, and second-order accurate nonstandard numerical method for solving the chemostat model. In the last section, a set of numerical simulations is presented that validate the theoretical findings.

2 A Simple Chemostat Model with Bacterial Input

The chemostat is an experimental device that was invented simultaneously by Monod and Novick-Szilard [6], which is widely used to model many ecological problems. In this paper, we examine a modification of the classical simple chemostat model [5, 6] by considering the dynamics of a single bacteria, B, and a growth-limiting substrate, S, under a constant input of the bacteria at the dilution rate D and bacterial cell death at the constant rate m. In this case, the modified chemostat model is given by the following system of ordinary differential equations:

$$\begin{aligned}\frac{dS}{dt} &= \underbrace{DS_{in}}_{\text{input}} - \underbrace{DS}_{\text{dilution}} - \underbrace{q\mu(S)B}_{\text{consumption by } B} = f_1(S, B),\\ \frac{dB}{dt} &= \underbrace{DB_{in}}_{\text{input}} - \underbrace{DB}_{\text{dilution}} + \underbrace{\mu(S)B}_{\text{growth}} - \underbrace{mB}_{\text{death of } B} = f_2(S, B),\end{aligned} \tag{1}$$

where S_{in} an B_{in} denote the concentrations of the input nutrient and bacterial biomass, respectively, and q is the yield constant. The growth rate function is given by the well-known Monod function:

$$\mu(S) = \frac{\mu^{\max} S}{K + S}, \tag{2}$$

where $\mu^{\max}$ is the maximal growth rate, and K is the half-saturation constant. A straightforward equilibrium analysis reveals that system (1) has one equilibrium $E^* = (S^*, B^*)$, where

$$B^* = \frac{DB_{in}}{\widehat{D}} + \frac{D}{q\widehat{D}}(S_{in} - S^*)$$

and

$$S^* = \begin{cases} \dfrac{S_{in} + \frac{\widehat{D}K}{\mu^{\max} - \widehat{D}} + \frac{q\mu^{\max} B_{in}}{\mu^{\max} - \widehat{D}} - \sqrt{\left(S_{in} + \frac{\widehat{D}K}{\mu^{\max} - \widehat{D}} + \frac{q\mu^{\max} B_{in}}{\mu^{\max} - \widehat{D}}\right)^2 - 4\frac{\widehat{D}S_{in}K}{\mu^{\max} - \widehat{D}}}}{2}, & \text{if } \widehat{D} < \mu^{\max} \\ \dfrac{\widehat{D}S_{in}K}{\widehat{D}K + q\mu^{\max} B_{in}}, & \text{if } \widehat{D} = \mu^{\max} \\ \dfrac{S_{in} + \frac{\widehat{D}K}{\mu^{\max} - \widehat{D}} + \frac{q\mu^{\max} B_{in}}{\mu^{\max} - \widehat{D}} + \sqrt{\left(S_{in} + \frac{\widehat{D}K}{\mu^{\max} - \widehat{D}} + \frac{q\mu^{\max} B_{in}}{\mu^{\max} - \widehat{D}}\right)^2 - 4\frac{\widehat{D}S_{in}k}{\mu^{\max} - \widehat{D}}}}{2}, & \text{if } \widehat{D} > \mu^{\max} \end{cases},$$

with $\widehat{D} = D + m$. In addition, one can also easily show that the cone

$$\mathbb{R}^2_+ = \{(S, B) : S, B \in \mathbb{R} \text{ and } S, B \geq 0\}$$

is positively invariant for system (1). Examining the Jacobian matrix

$$J(S, B) = \begin{pmatrix} -D - qB\frac{\mu^{\max}K}{(K+S)^2} & -q\mu(S) \\ B\frac{\mu^{\max}K}{(K+S)^2} & \mu(S) - \widehat{D} \end{pmatrix}, \tag{3}$$

at the equilibrium $E^* = (S^*, B^*)$, one can show that $E^* = (S^*, B^*)$ is locally asymptotically stable, since $det(J(S_i^*, B_i^*)) > 0$ and $-trace(J(S_i^*, B_i^*)) > 0$. Finally, using the Dulac's Criterion and applying the Poincaré-Bendixson's Theorem we can conclude that E^* attracts any trajectory in $\triangle = \{(S, B) : S + qB \leq S_{in} + qB_{in}\}$, and therefore, E^* is also globally asymptotically stable.

3 A Higher-Order NSFD Method

In this section, we present the new second-order accurate, positivity-preserving, and elementary stable nonstandard (SOPESN) numerical method for system (1):

$$\begin{aligned} \frac{S_{k+1} - S_k}{\varphi_1(h, S_k, B_k)} &= w_1^k\big(D(S_{in} - S_k) - q\mu(S_k)B_k\big), \\ \frac{B_{k+1} - B_k}{\varphi_2(h, S_k, B_k)} &= w_2^k\big(D(B_{in} - B_k) + \mu(S_k)B_k - mB_k\big). \end{aligned} \tag{4}$$

where

$$w_1^k = \begin{cases} 1, & \text{if } f_1(S_k, B_k) \geq 0 \\ \frac{S_{k+1}}{S_k}, & \text{if } f_1(S_k, B_k) < 0 \end{cases} \quad \text{and} \quad w_2^k = \begin{cases} 1, & \text{if } f_2(S_k, B_k) \geq 0 \\ \frac{B_{k+1}}{B_k}, & \text{if } f_2(S_k, B_k) < 0 \end{cases}.$$

Here, (S_k, B_k) denotes the approximation of the exact solution $(S(t_k), B(t_k))$ to the system (1), where $t_k = kh$, with k positive integer and step-size $h > 0$. The modified nonstandard denominator function $\varphi_i : \mathbb{R}_+ \times \mathbb{R}^2_+ \to \mathbb{R}_+$ in the proposed SOPESN method is chosen as follows:

$$\varphi_i(h, S, B) = h - q_i(S, B)\frac{h^2}{2} + O(h^3), \tag{5}$$

where

$$q_i(S, B) = \begin{cases} -\Big(\dfrac{\partial f_i(S, B)}{\partial S}\dfrac{f_1(S, B)}{f_i(S, B)} + \dfrac{\partial f_i(S, B)}{\partial B}\dfrac{f_2(S, B)}{f_i(S, B)}\Big), & f_i(S, B) > 0 \\ \dfrac{2f_i(S, B)}{X_i} - \Big(\dfrac{\partial f_i(S, B)}{\partial S}\dfrac{f_1(S, B)}{f_i(S, B)} + \dfrac{\partial f_i(S, B)}{\partial B}\dfrac{f_2(S, B)}{f_i(S, B)}\Big), & f_i(S, B) < 0 \end{cases},$$

with $X_1 = S$ and $X_2 = B$.

Theorem 1 *The NSFD method* (4) *for approximating the solutions of the continuous model* (1) *is positivity-preserving, elementary stable, and of second order accurate when using the modified nonstandard denominator function* (5).

Proof Positivity of the method can be seen by rewriting system (4) in the following explicit form:

$$\begin{aligned} S_{k+1} &= \begin{cases} S_k + \varphi_1(h, S_k, B_k) f_1(S_k, B_k), & \text{if } f_1(S_k, B_k) \geq 0 \\ \dfrac{S_k^2}{S_k - \varphi_1(h, S_k, B_k) f_1(S_k, B_k)}, & \text{if } f_1(S_k, B_k) < 0 \end{cases}, \\ B_{k+1} &= \begin{cases} B_k + \varphi_2(h, S_k, B_k) f_2(S_k, B_k), & \text{if } f_2(S_k, B_k) \geq 0 \\ \dfrac{B_k^2}{B_k - \varphi_2(h, S_k, B_k) f_2(S_k, B_k)}, & \text{if } f_2(S_k, B_k) < 0 \end{cases}. \end{aligned} \tag{6}$$

Since $\varphi_i(h, S_k, B_k) > 0$, then clearly $S_k > 0$, implies $S_{k+1} > 0$, and $B_k > 0$ implies $B_{k+1} > 0$.

Next, it is easy to see from the formulation (4) that the equilibrium $E^* = (S^*, B^*)$ of Equation (1) is a fixed point of the NSFD method and vice versa, and that there are no other fixed point of (4). Finally, since $E^* = (S^*, B^*)$ is locally asymptotically stable, the eigenvalues of the Jacobian matrix (3) evaluated at $E^* = (S^*, B^*)$ are $\lambda_i < 0$ for all $i = 1, 2$. To show elementary stability we next consider the linearized version of system (1):

$$(S'(t), B'(t))^T = J(S^*, B^*)(S(t) - S^*, B(t) - B^*)^T. \tag{7}$$

Since the matrix $J(S^*, B^*)$ is diagonalizable, there is a 2×2 invertible matrix P such that $\text{diag}(\lambda_1, \lambda_2) = P^{-1}J(S^*, B^*)P$. Using the change of variable $(\bar{S}, \bar{B})^T = P^{-1}(S(t) - S^*, B(t) - B^*)^T$, then it can be easily seen that Eq. (7) is equivalent to

$$(\bar{S}'(t), \bar{B}'(t))^T = diag(\lambda_1, \lambda_2)(\bar{S}, \bar{B})^T. \tag{8}$$

After applying the SOPESN method to Eq. (8) and since $\lambda_i < 0$ for all $i = 1, 2$, then

$$\bar{S}_{k+1} = \frac{\bar{S}_k}{1 - \varphi_1(h, \bar{S}_k, \bar{B}_k)\lambda_1}, \bar{B}_{k+1} \qquad = \frac{\bar{B}_k}{1 - \varphi_2(h, \bar{S}_k, \bar{B}_k)\lambda_2}. \tag{9}$$

Since the eigenvalues $\lambda_i < 0, \forall i = 1, 2$, then clearly $1 - \varphi_i(h, \bar{S}_k, \bar{B}_k)\lambda_i > 1$ and hence $0 < \frac{1}{1-\varphi_i(h, y_1^k, y_2^k)\lambda_1} < 1$ which implies $\bar{S}_i^k \to 0$ and $\bar{B}_k \to 0$ as $k \to \infty$, i.e., $E^* = (S^*, B^*)$ is a stable fixed point of the SOPESN method (4).

Finally, to prove the second-order accuracy of the SOPESN method, let us first consider the case of $f_i(S, B) > 0$. Using Taylor series expansion about t_k yields

$$\begin{aligned} S(t_{k+1}) - \Big[S(t_k) + \varphi_1(h, S(t_k), B(t_k)) f_1(S(t_k), B(t_k))\Big] &= f_1(S(t_k), B(t_k))h \\ + \Big(\frac{\partial f_1(S(t_k), B(t_k))}{\partial S} f_1(S(t_k), B(t_k)) + \frac{\partial f_1(S(t_k), B(t_k))}{\partial B} f_2(S(t_k), B(t_k))\Big)\frac{h^2}{2} \\ - \varphi_1(h, S(t_k), B(t_k)) f_1(S(t_k), B(t_k)) + O(h^3) = O(h^3), \end{aligned}$$

and, similarly,

$$B(t_{k+1}) - \Big[B(t_k) + \varphi_2(h, S(t_k), B(t_k)) f_2(S(t_k), B(t_k))\Big] = O(h^3).$$

Similarly, one can show the second-order accuracy of the new NSFD method in the case of $f_i(S, B) < 0$, by using Taylor series expansion about t_k and also Maclaurin series of

$$\frac{1}{1 - \frac{f_i(S,B)\varphi_i(h,S,B)}{X_i}}.$$

4 Numerical Simulations and a Conclusion

To illustrate our theoretical results we use the new SOPESN method (4) with the following nonstandard denominator function

$$\varphi_i(h, S_k, B_k) = \frac{1 - e^{-q_i(S_k, B_k)h}}{q_i(S_k, B_k)}, \quad i = 1, 2, \tag{10}$$

where $q_i(S, B)$ is selected as in Sect. 3 which satisfies the conditions of Theorem 1, and compare it to the second-order explicit Runge-Kutta (ERK2) method, the Explicit Euler (EE) method, and the first-order positivity-preserving and elementary stable nonstandard (PESN) method [7].

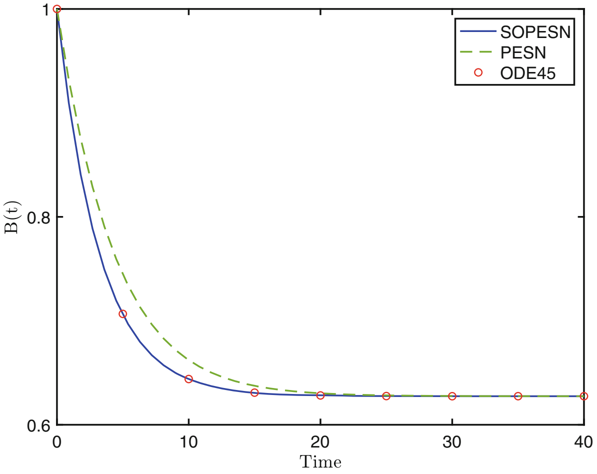


Fig. 1 Comparison of SOPESN and PESN method when $\mu_{\max} = 0.3$, $m = 0.2$, $D = 0.4$, $K = 0.1$, $q = 10^{-8}$, $B_{in} = 0.5$, $S_{in} = 1.5$ and $h = 0.9$

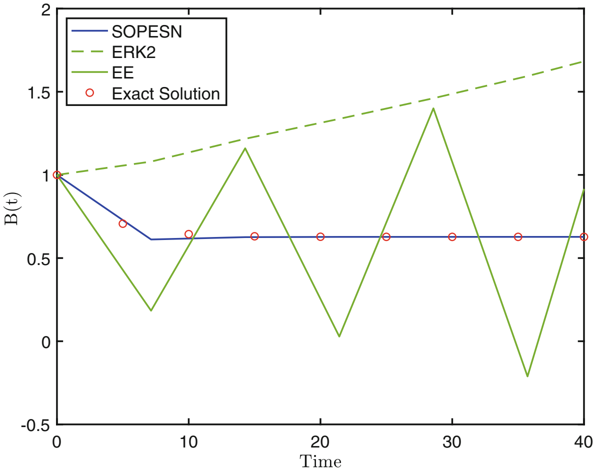


Fig. 2 Comparison of SOPESN, ERK2 and EE numerical methods when solving the model (1), using $\mu^{\max} = 0.3$, $m = 0.2$, $D = 0.4$, $K = 0.1$, $q = 10^{-8}$, $B_{in} = 0.5$, $S_{in} = 1.5$ and $h = 7$

In Fig. 1, we compare our SOPESN method with the first-order PESN method, for $h = 0.9$ and initial conditions $S(0) = 2$, $B(0) = 1$, and can see that the SOPESN method converges much faster to the exact solution. As can be seen in Fig. 2, for $h = 7$ and initial conditions $S(0) = 2$, $B(0) = 1$, the numerical solution from the ERK2 method increases and eventually blows up to infinity as time increases, i.e., does not preserve the stability property of the equilibrium. Similarly, the EE method does not preserve the positivity and elementary stability either, as the numerical solution oscillates before eventually blowing up. However, the SOPESN method's numerical solution converges to the exact solution, for any size of the time step $h > 0$, with a second-order accuracy while preserving the positivity and local stability property of the equilibrium.

The numerical simulations presented here demonstrate the advantages of the new SOPESN method as compared to standard numerical methods, which shows the importance of the elementary stability and positivity-preserving properties of the new NSFD method in addition to its higher-order accuracy than traditional NSFD methods.

References

1. Anguelov, R., and J. M.-S. Lubuma. “Contributions to the mathematics of the nonstandard finite difference method and applications.” Numer. Meth. Partial Diff. Eqs. 17, no. 5 (2001): 518–543.
2. Dimitrov, D.T., and H.V. Kojouharov. “Nonstandard Numerical Methods for a Class of Predator-Prey Models with Predator Interference.” Elec. J. Diff. Eqs. 15 (2007): 67–75.
3. Gupta, M., J.M. Slezak, F. Alalhareth, S. Roy, and H.V. Kojouharov. “Second-order nonstandard explicit Euler method.” AIP Conference Proceedings 2302, no. 1 (2020): 110003.
4. Kojouharov, H.V., S. Roy, M. Gupta, F. Alalhareth, and J.M. Slezak. “A second-order modified nonstandard theta method for one-dimensional autonomous differential equations.” Appl. Math. Lett. 112 (2021): 106775.
5. Martines, I.P., H.V. Kojouharov, J.P. Grover. “A chemostat model of resource competition and allelopathy”. Appl. Math. Comput. 215 (2009): 573–582.
6. Smith, H.L., and P. Waltman. The theory of the chemostat: dynamics of microbial competition. Vol. 13. Cambridge University Press, 1995.
7. Wood, D.T., H.V. Kojouharov. “A class of nonstandard numerical methods for autonomous dynamical systems.” Appl. Math. Lett. 50 (2015): 78–82.

Stability and Convergence of a Class of RKDG Methods for Maxwell's Equations

Adérito Araújo and Sunčica Sakić

Abstract This paper is concerned with a Runge-Kutta Discontinuous Galerkin (RKDG) method for solving the time-dependent Maxwell's equations in the context of light propagation in the human eye. The method is based on a discontinuous Galerkin (dG) method for the spatial discretisation of the partial differential equations (PDEs) and an explicit fourth-order Runge-Kutta method for the integration of the resulting system of ordinary differential equations (ODEs). The stability and convergence properties of the method are studied and experimentally inspected.

1 Introduction

In recent decades, medical imaging techniques have contributed significantly to the understanding of the internal structures of the eye, as well as to the diagnosis and treatment of many diseases. Our research group is particularly interested in investigating efficient algorithms to simulate the propagation of light in the human eye, with the aim of identifying the conditions that lead to different pathologies [1, 2]. In [3], we used Maxwell's equations to model the electromagnetic wave propagation through the cornea, in order to explain which factors lead to the deterioration of corneal transparency.

Maxwell's equations are the fundamental set of equations that describe the behaviour of an electromagnetic wave interaction with materials [7]. In this work, we consider the time-dependent Maxwell's equations in the transverse electric (TE) mode [8] that correspond, in the conservative form, to the two dimensional system of PDEs

A. Araújo (✉)
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: alma@mat.uc.pt

S. Sakić
Department of Numerical Mathematics, Charles University, Prague, Czech Republic
e-mail: sakic@karlin.mff.cuni.cz

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_64

$$\begin{pmatrix} \varepsilon & 0 \\ 0 & \mu \end{pmatrix} \frac{\partial u}{\partial t} + \nabla \cdot \begin{pmatrix} 0 & -H_z \\ H_z & 0 \\ E_y & -E_x \end{pmatrix} = 0, \qquad (x, y, t) \in \Omega \times (0, T], \tag{1}$$

with $u = (E_x, E_y, H_z)^\top$, where E_x and E_y are electric field components and H_z is the magnetic field component, respectively. Here Ω is a bounded polygonal domain in $\mathbb{R}^2$, T is a real scalar, $\nabla\cdot$ the divergence operator, and ε and μ are the electric permittivity and the magnetic permeability of the medium, respectively. The electric permittivity tensor ε is two dimensional symmetric, uniformly positive definite (for almost every $(x, y) \in \Omega$) and uniformly bounded with a strictly positive lower bound, while the magnetic permeability μ is a scalar function varying in space. The model is completed with initial conditions and perfect conductor boundary conditions [9].

2 Discretisation Method

In this section, we introduce the numerical method for the discretisation of (1). We will consider a method of lines approach to the numerical solution: we first discretise the PDEs in space with a nodal dG method [6] and then we integrate the resulting system of ODEs with an explicit Runge-Kutta type method [4].

For the space discretisation, we consider a conformal triangulation $\mathcal{T}_h = \{T_k\}_{k=1}^K$ of Ω, with a spatial discretisation parameter $h = \max_{k=1,\dots,K} \operatorname{diam}(T_k)$. On each triangle T_k, the solution is approximated by a polynomial u_h^k of degree less than or equal to N. The global solution is approximated by a numerical solution u_h obtained by a direct sum of the K local polynomial solutions u_h^k, connecting all local solutions via numerical fluxes [2]. The space discretisation leads to a linear system of ODEs

$$\frac{du_h}{dt} = \mathcal{L}_h(\alpha) u_h, \tag{2}$$

where the parameter α is related to the numerical flux used in the discretisation: for $\alpha = 0$, we have a central flux, and for $\alpha = 1$, we have an upwind flux. For a fixed mesh, it is clear that operator $\mathcal{L}_h$ depends only on α.

To define the fully discrete scheme, the system of ODEs (2) needs to be integrated in time. A common choice [6] is to use the standard explicit 4-stage fourth order Runge-Kutta (ERK4) method. The major drawback is that this method is memory demanding and has a small stability region. To overcome these limitations, we will consider an explicit low-storage Runge-Kutta type method. One of the best known methods of this class is the low-storage 5-stage fourth-order LSERK(5,4) method, introduced in [5]. When compared with the standard ERK4, this method reduces memory requirements without significantly increasing computational costs.

An alternative to LSERK(5,4) is the version that uses 14 stages, proposed by Niegemann et al. in [10], the LSERK(14,4). Unlike the five stage version, this method come with extra costs due to the more internal stages. However, the advantage of LSERK(14,4) reflects in the stability of the RKDG method as we will see in the next section.

3 Stability and Convergence

In this section we analyse the stability and convergence of the proposed RKDG method for the normalised TE-mode of Maxwell's equations, obtained by making dimensionless each quantity that appears in initial equations. Informally speaking, we consider $\varepsilon = \mathrm{diag}(1, 1)$ and $\mu = 1$. All experiments shall be performed on domain $\Omega = (-1, 1)^2$ tessellated into structured grids of various sizes (see Table 1).

3.1 Stability

To analyse the stability of the RKDG method, we start by illustrating the behaviour of the spectrum of the semi-discrete dG operator $\mathcal{L}_h$ given in (2). In our experiments, we consider the mesh parameters from Table 1 for $K = 32$ elements and we fix the order of the polynomial approximating the local solution at $N = 4$. The number of interpolation points on each triangle T_k is $N_p = (N+1)(N+2)/2 = 15$. Since there are three fields whose solution is unknown (E_x, E_y and H_z), the dG semi-discrete operator is of size $K \times N_p \times 3 = 1440$.

Each of plots in Fig. 1 corresponds to 1440 eigenvalues λ of the semi-discrete dG operator $\mathcal{L}_h$ given in (2), denoted by blue asterisks, and $\Re(\lambda)$) and $\Im(\lambda)$ corresponds to the real and imaginary part of the eigenvalue λ, respectively. Note that, by decreasing the parameter α, the spectrum is getting closer to the imaginary axis. Consequently, for $\alpha = 0$ all eigenvalues become purely imaginary. This is a result of the energy-conserving nature of central flux.

In Fig. 2 the regions of absolute stability of the three proposed numerical integrators (ERK4, LSERK(5,4) and LSERK(14,4)) are illustrated and compared with the eigenvalues λ of the semi-discrete operator for the upwind case. As we may see, LSERK(14,4) has the widest stability region. In other words, LSERK(14,4) allows bigger time-steps Δt to integrate (2) without compromising its stability properties.

Table 1 Description of meshes used for numerical tests

Minimal distance between two vertices ($h_{\min}$)	0.70	0.56	0.28	0.14	0.07
Number of triangles (K)	32	50	200	800	3200
Number of vertices (N_v)	25	36	121	441	1681

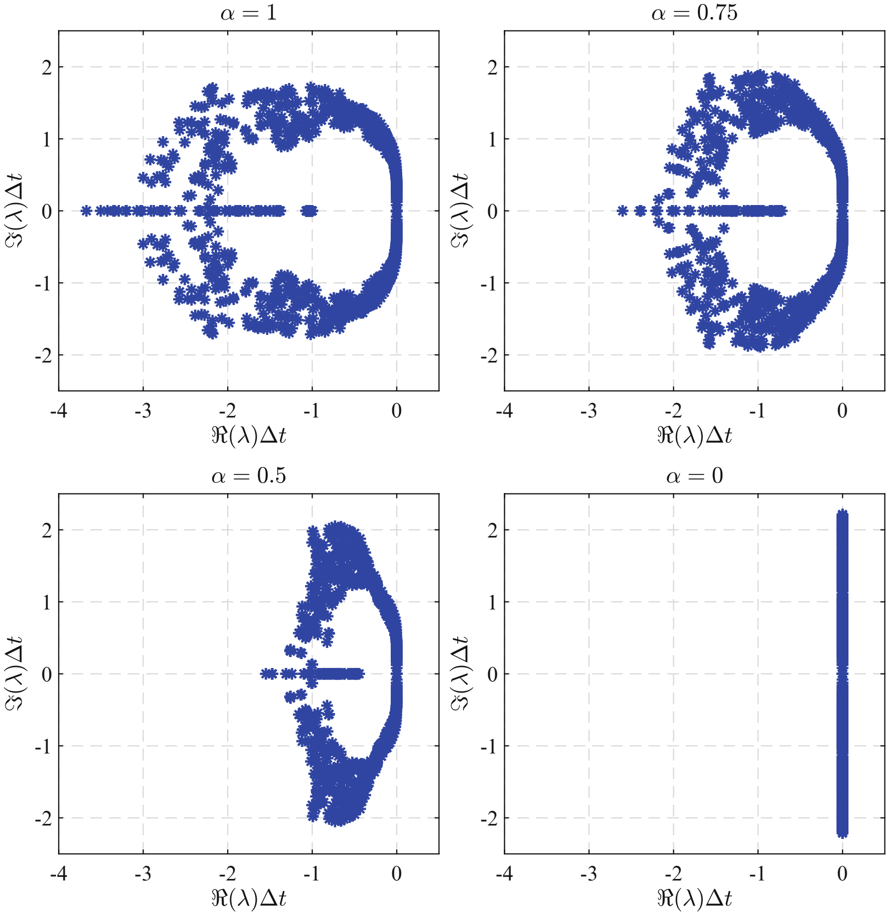


Fig. 1 Spectrum of dG operator for TE-mode of Maxwell's equations

3.2 Convergence

For purposes of convergence analysis, we set the initial condition as the planar wave $(E_x(x, y, 0), E_y(x, y, 0), H_z(x, y, 0))^\top = (0, 0, \cos(\pi x)\cos(\pi y))^\top$. The exact solution of (1) with perfect conductor boundary may be easily constructed. In our numerical tests, we used the LSERK(14,4) method and the final time was set to be $T = 0.1$. The difference between the exact and approximate solutions in L^2-norm are computed, and the associated orders of convergence in space and time are determined.

To illustrate the order of convergence in space, we fixed the time-step at $\Delta t = 10^{-4}$. The computations are performed for both central and upwind fluxes on different meshes, given in Table 1, and the polynomial degrees varied from one to

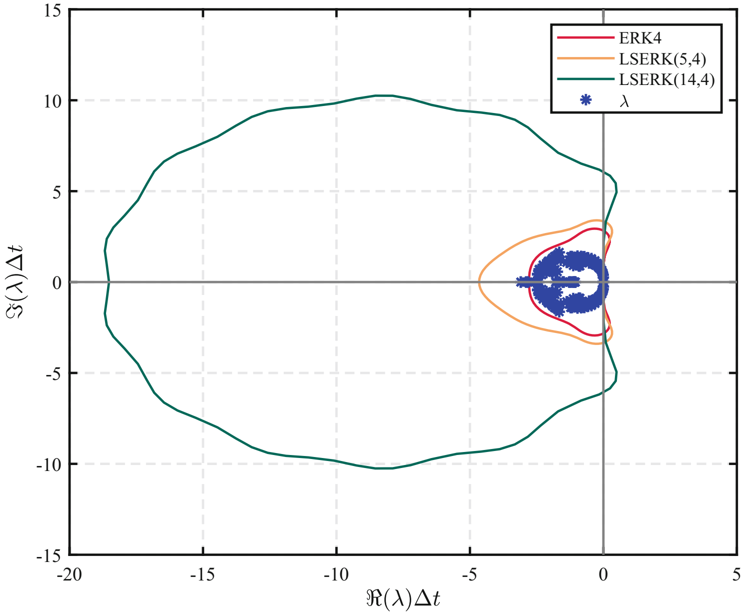


Fig. 2 Spectrum of dG operator in stability regions of ERK methods

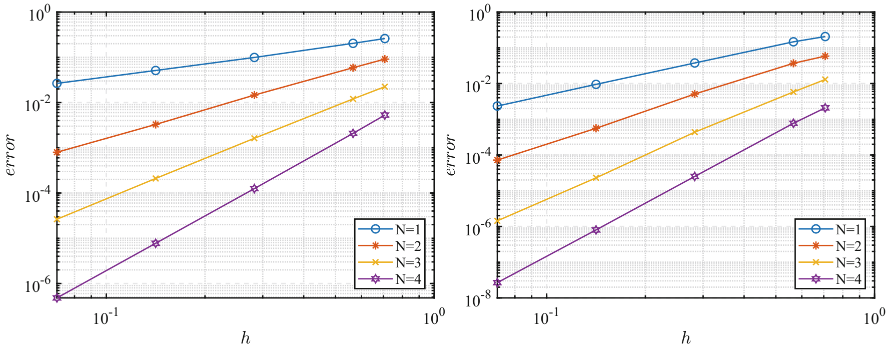


Fig. 3 L^2-error for field E_x *versus* h for central flux (left) and upwind flux (right)

four. In Fig. 3 we present the L^2-errors for the component E_x of electric field, into a plot whose axes are logarithmically scaled. As we may see, when the central flux is used, the order of convergence is around $\mathcal{O}(h^N)$, while for upwind fluxes we observe higher order, up to $\mathcal{O}(h^{N+1})$, in accordance with theoretical results presented in the literature (see [1] and the references therein).

To visualise the convergence in time, the polynomial degree and the number of elements in the mesh have been set to $N = 8$ and $K = 3200$, respectively. In Fig. 4 the L^2-errors for the component E_x of electric field is computed while decreasing

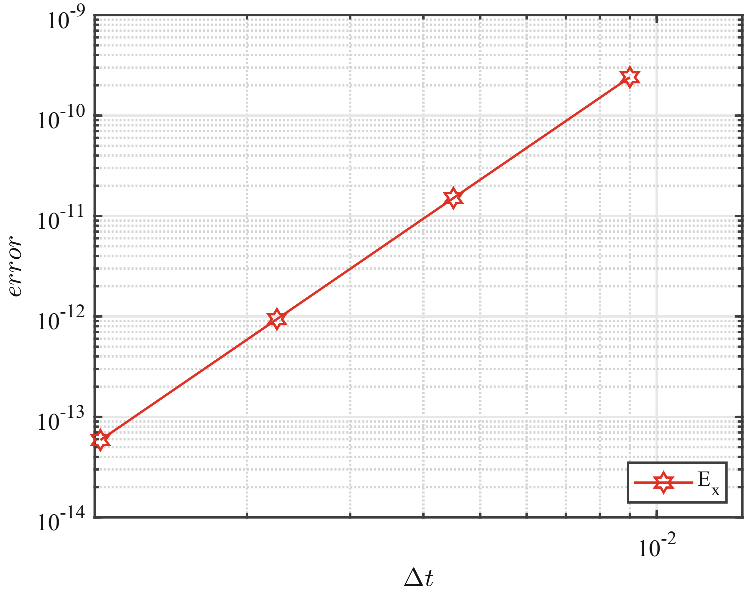


Fig. 4 L^2-error for field E_x *versus* Δt

the time-step Δt. The results illustrate the fourth-order convergence in time. Same results were obtained for both central and upwind fluxes.

4 Conclusion

We studied the stability and convergence of a class of RKDG methods for the discretisation of the time-domain TE-mode Maxwell's equations. This class uses a dG method for the spatial discretisation and a fourth-order Runge-Kutta type integrator for solving the resulting system of ODEs. When the LSERK(14,4) time integrator is consider, it was shown that the spectra of the fully discrete scheme remain in its stability region. The convergence of the method was analysed in both space and time, confirming the theoretical convergence rates for both upwind and central fluxes.

Acknowledgments This work was partially supported by the Centre for Mathematics of the University of Coimbra (UIDB/00324/2020, funded by the Portuguese Government through FCT/MCTES) and co-funded by FEDER—Fundo Europeu de Desenvolvimento Regional, through COMPETE 2020—Programa Operacional Fatores de Competitividade, and the National Funds through FCT—Fundação para a Ciência e a Tecnologia, Project POCI-01-0145-FEDER-028118, PTDC/MAT-APL/28118/2017.

References

1. Araújo, A., Barbeiro, S., Galati, M. Kh.: Stability of a leap-frog discontinuous Galerkin method for time-domain Maxwell's equations in anisotropic materials. Commun. Comput. Phys. **21**(5), 1350–1375 (2017)
2. Araújo, A., Barbeiro, S., Galati, M. Kh.: Convergence of an explicit leap-frog discontinuous Galerkin method for time-domain Maxwell's equations in anisotropic materials. J. Ind. Math. **8**(9), 1350–1375 (2018)
3. Araújo, A., Barbeiro, Bernardes, R., Morgado, M., Sakić, S: A mathematical model for the corneal transparency problem. submitted.
4. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM (1998)
5. Carpenter, M.H., Kennedy, C.A.: Fourth-order 2N-storage Runge-Kutta schemes. NASA Report TM 109112, NASA Lengley Research Center (1994)
6. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer Verlag, New York (2008)
7. Jin, J.M.: Theory and Computation of Electromagnetic Fields, 1st edn. Wiley-IEEE Press (2010)
8. König, M., Busch, K., Niegemann, J.: The Discontinuous Galerkin Time-Domain method for Maxwell's equations with anisotropic materials. Photonics Nanostructures: Fundam. Appl. **8**, 303–309 (2010)
9. Niegemann, J., König, M., Stannigel, K., Busch, K.: Higher-order time-domain methods for the analysis of nano-photonic systems. Photonics Nanostructures: Fundam. Appl. **7**(1), 2–11 (2009)
10. Niegemann, J., Diehl, R., Busch K.: Efficient low-storage Runge-Kutta schemes with optimized stability regions. J. Comput. Phys. **231**, 364–372 (2012)

Safeguarding the Nation's Digital Memory: Bayesian Network Modelling of Digital Preservation Risks

Martine J. Barons, Thais C. O. Fonseca, Hannah Merwood, and David H. Underdown

Abstract Archives comprise primary sources which may be physical, born digital or digitised. Digital records have a limited lifespan, through carrier degradation, software and hardware obsolescence and storage frailties. It is important that the original bitstream of these primary sources is preserved and can be demonstrated to have been preserved. Soft elicitation with experienced archivists was used to identify the most likely elements contributing to digital preservation success and failure and the relationships between these elements. A Bayesian Network representation of an integrating decision support system provided a compact representation of reality, enabling the risk scores for various scenarios to be compared using a linear utility function. Thus, the effect on risk of various actions and interventions can be quantified. This tool, DiAGRAM, is now in use.

1 Introduction

Archives comprise primary sources which can be physical, born digital and digitised. Digital records have a limited lifespan, through carrier degradation, software and hardware obsolescence and storage frailties. It is important that the original bitstream of these primary sources is preserved and can be demonstrated to

M. J. Barons (✉)
AS&RU, Department of Statistics, University of Warwick, Coventry, UK
e-mail: martine.barons@warwick.ac.uk

T. C. O. Fonseca
Department of Statistical Methods, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: thais@im.ufrj.br

H. Merwood
Government Operational Research Service, London, UK

D. H. Underdown
The National Archives, Surrey, UK
e-mail: david.underdown@nationalarchives.gov.uk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_65

have been preserved; this consumes significant resources [1]. Digital preservation (DP) is crucial for ensuring the longevity of societal history, for research, legal accountability, government and business planning. It is a maturing field, with the main standards around 20 years old. Larger (and relatively better funded) archives such as national, state and provincial archives in high income countries have been engaged in DP for longer periods. It is now becoming a pressing issue for all archives.

The archival sector typically lacks sufficient people, sufficiently skilled people and sufficient funding to undertake all possible mitigations against these risks. Thus, there is a need for support in choosing the mitigation strategies which bring the largest and most immediate reduction in overall risk levels in the current context of an individual archive: there is not a ‘one-size fits all’ solution.

The National Archives in the United Kingdom (TNA), and the Applied Statistics & Risk Unit (AS&RU) at the University of Warwick collaborated to build decision support suitable for identifying risks to digital archives and quantifying the efficacy of mitigation strategies, the Digital Archiving Graphical Risk Assessment Model (DiAGRAM) [2].

2 Methodology

Soft elicitation [3] with experienced archivists was used to identify the most likely elements contributing to digital preservation success and failure and the relationships between these elements. This established, it became obvious that a Bayesian Network [4, 5] representation of an integrating decision support system (IDSS, [6, 7] would be appropriate as a compact representation of reality in this case. However, not all the data required to quantify the model was available, so structured expert judgement was employed to provide data in the gaps. The IDSS is a new paradigm for drawing together evidence from different parts of large systems to provide decision support. Each part of the system is typically overseen by a panel of domain experts using their own data and, often complex, models. Panels contribute key summaries of future expectations under different candidate policy decisions. The IDSS then allows the decision centre to calculate expected utility scores for these candidate policies for comparison and decision support.

2.1 Bayesian Networks

A discrete Bayesian Network as defined in [4] is a compact representation of the joint probability distribution $p(\mathbf{x})$ of a p-variate vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)'$. The model is specified by the set $\mathcal{N} = (\mathcal{X}, \mathcal{G}, \mathcal{P})$ with elements given by

1. a graph $\mathcal{G} = (V, E)$ with nodes V and connections E;
2. a set o variables $\mathcal{X}$ representing the nodes of $\mathcal{G}$;
3. a set of conditional distributions $\mathcal{P}$ with distribution $p_i(x_i \mid x_{pa(i)})$ for each $X_i \in \mathcal{X}$,

where $X_{pa(i)}$ is the set of parents of X_i. A Bayesian Network model is composed by the representation induced by $\mathcal{N}$ which is given by

$$p(\mathbf{x}) = \prod_{v \in \mathcal{X}} p_i(x_v \mid x_{pa(v)}).$$

The inferential problem depends on the computation of $P(X_v = x_v \mid \epsilon)$, $X_v \in \mathcal{X}$ given a set of evidences ϵ, that is, the computation of total probabilities depending on sums and multiplications. However, this computation is costly even for small p. Often algorithms such as Logic Sampling are used to approximate the predictive probabilities of interest. In the context of categorical data, the distributions assumed for the observations are multinomial such that $X_i \mid X_{pa(i)} = j, \boldsymbol{\theta}_{ij} \sim Mult(M_{ij}, \boldsymbol{\theta}_{ij})$ and are represented as conditional probability tables (CPTs). If a Dirichlet prior with parameter $\boldsymbol{a}_{ij}$ is assumed for $\boldsymbol{\theta}_{ij}$ then the posterior distribution is Dirichlet with parameter $\boldsymbol{N}_{ij} + \boldsymbol{a}_{ij}$ with N_{ijk} the counts of $\{X_{ik} = x_{ik}\}$ when $\{X_{pa(i)} = j\}$. For a practical guide on how to perform inference and prediction using Bayesian Networks see [8].

Where data was not available, Structured Expert Judgement (SEJ) was used to quantify experts' uncertainties on the values for the conditional probability tables.

2.2 *Structured Expert Judgement*

Expert judgement is pervasive in all forms of risk analysis [7]. Structured expert judgement elicitation is a well-established paradigm for eliciting expert judgements of uncertain quantities and event occurrences [9]. Structured protocols seek to mitigate the most pervasive cognitive frailties when asking for subjective judgements, such as group-think, availability bias, personality effects and overconfidence. We used the recently-developed IDEA protocol [10]. Calibration questions are included, drawn from existing surveys and reports, on which individual experts' accuracy and informativeness can be calculated for performance-weighted pooling of the results into a single distribution, using the classical approach [11].

3 DiAGRAM

The Digital Archiving Graphical Risk Assessment Model (DiAGRAM) is a bespoke tool developed to facilitate the computation of digital preservation risks and provide comparison of competing policies. It aims to improve users' understanding of digital

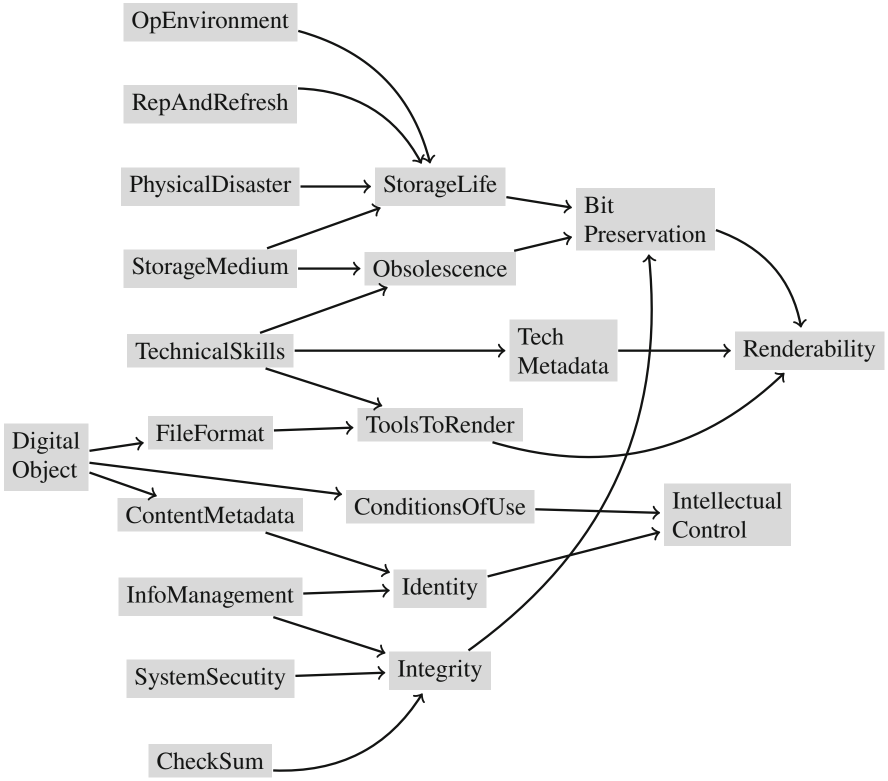


Fig. 1 Qualitative description of the digital preservation system

archiving risks, empower archivists to compare and prioritise different threats to the digital objects and to aid in quantifying the impact of risk events and risk management strategies on digital preservation to support decision making.

The model contains the network elicited using soft elicitation $\mathcal{G}$ and the conditional probability tables $\mathcal{P}$ representing the uncertainty in the nodes obtained via historical data when it is available, and through SEJ elicitation otherwise.

3.1 Network Structure Construction

The variables and the qualitative relationships between them were elicited through close communication with domain experts. The experts' collective views were represented by a Directed Acyclic Graph (DAG) (Fig. 1). The variables included in the model were Digital Object, Identity, Conditions of Use, Intellectual Control, Information Management, Technical Skills, Operating Environment, Content Metadata, Technical Metadata, Checksum, File Format, Bit-preservation, Obsolescence, Tools to render, Storage Medium, Storage Life, Replication and Refreshment, System Security, Integrity and Renderability.[1] The variables Intellectual Control and Renderability comprise the utility function which provides comparative scores for candidate policies in the policy comparison step. See [2] for further details on structure construction and node definitions.

3.2 *Expert Elicitation Results*

SEJ was used in DiAGRAM to quantify Storage life, Obsolescence, Technical Metadata, Tools to Render, Conditions of Use, Content Metadata, Identity, Integrity, Bit Preservation and Renderability. In the elicitation sessions, 22 participating experts answered 20 calibration questions and 24 questions of interest. The transformed Kullback-Leibler divergence and the performance-weighted outcomes for all experts are presented in Table 1. The results show experts 8, 12 and 16 had the best performances on the calibration questions and experts 13, 20, and 21 had the worst performances.

3.3 *Joint Probability Distribution*

This section computes the probability distributions based on the structure, tables elicited from experts and data available. The data sources used were: the 2019 JISC digital skills survey of over 300 UK archive professionals; the cloud data storage providers on access and durability; data from the Environment Agency on the long-term flood risk of UK postcodes; and data from TNA on file formats by digital object type.

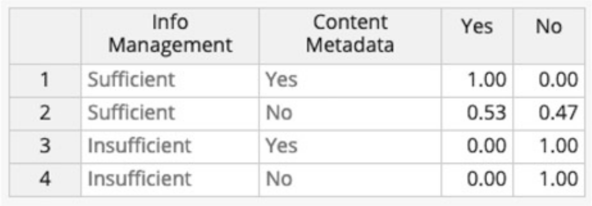
In DiAGRAM, of the 21 nodes, 9 have the probabilities customisable by the users to reflect their institution: Digital Object, Operating Environment, Replication and Refreshment Storage Medium, Technical Skills, Information Management, System Security and Checksum.

For comparative purposes DiAGRAM provides a Baseline Model (BM) where the customisable nodes are set to: (1) no technical skills; (2) good level of system security (74%); (3) 0% of files have a check-sum; (4) 14% of files have sufficient internal information management systems in place; (5) 100% of the digital archive is born digital; (6) 100% of storage media are stored on outsourced (cloud) storage; (7) 100% of files have a good replication and refreshment strategy in place; (8) operating environment was considered 100 % as all files have copies in different locations; (9) The risk of physical disaster (flood risk rating) is very low. For this baseline model, the conditional probability table for the node Identity obtained in

[1] See DiAGRAM's 'Glossary' tab here: https://nationalarchives.shinyapps.io/DiAGRAM/.

Table 1 Experts ID, experts transformed Kullback-Leibler (KL) divergence and final combined weights (×10^3). Weights close to 0 indicate worse performances and close to 1 (×10^3) indicate better performances

Expert ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
KL divergence	14	11	10	25	20	9	36	6	37	25	17	6	45	19	19	6	24	26	28	48	66	8
Final weight	7	25	33	0	0	38	0	229	0	0	3	209	0	1	1	295	0	0	0	0	0	158



	Info Management	Content Metadata	Yes	No
1	Sufficient	Yes	1.00	0.00
2	Sufficient	No	0.53	0.47
3	Insufficient	Yes	0.00	1.00
4	Insufficient	No	0.00	1.00

Fig. 2 Conditional probability table for the node Identity obtained in DiAGRAM for the baseline Commercial Backup model

DiAGRAM for this setup is presented in Fig. 2. Probability tables for all nodes can be obtained and are used to compute the final utility function.

3.4 Utility Computation and Scenario Evaluation

In consultation with a wide range of digital archivists, the utility for DiAGRAM was defined as Renderability and Intellectual Control. Renderability (R) captures the need for the digital object to have a sufficiently useful representation of the original file. 'Sufficiently useful' depends on the use to which a digital object is being put. Intellectual Control (IC) is the archivist's need to have full knowledge of the digital object's content, provenance and conditions of use. IC requires sufficient metadata that the archivist can identify the appropriate object, see how it relates to other objects from the same source, and understand whether they have the copyright permissions to make reproductions, or if data protection, etc. prevents the object from being made publicly available (and how long those restrictions will remain applicable).

We compare the Baseline Model with the alternative scenario of Commercial Backup (CB), which is as for BM but improving information management to 43 % and technical skill level to 30 %. The risk scores for BM and the CB are compared using a linear utility function (Fig. 3). The CB scenario has larger total score (62: IC = 20, R = 42) than BM (44: IC = 6, R = 38), showing that moving to CB improves digital preservation.

Acknowledgments The authors acknowledge with gratitude valuable contributions to the project by staff from partner institutions,[2] and by the additional experts who participated in the elicitation. This work was supported by: the National Lottery Heritage Fund under project reference number OM-19-01060; The Engineering and Physical Sciences Research Council under grant EP/R511808/1; and The National Archives (UK).

[2] See project webpage: https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/.

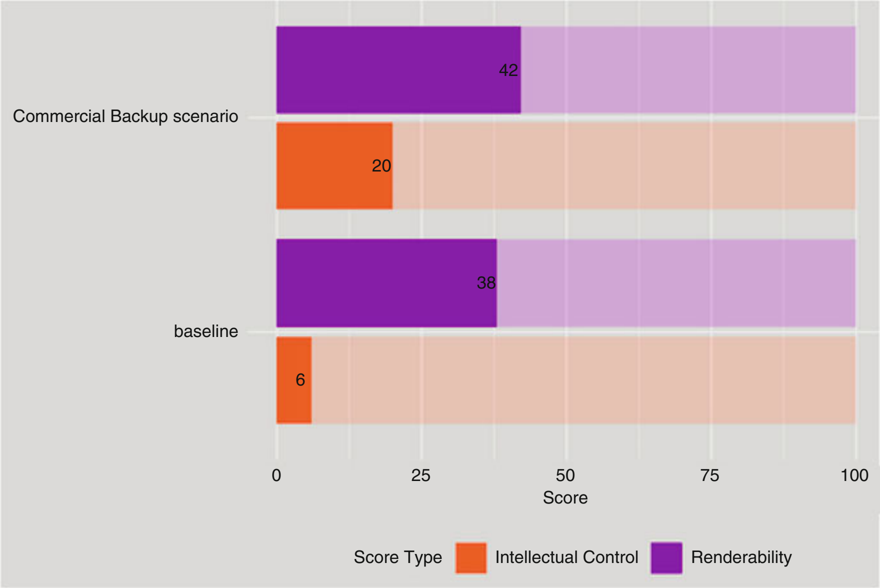


Fig. 3 Intellectual control and renderability scores comparison for the baseline commercial backup model and the commercial backup scenario

References

1. R.D. Frank. The social construction of risk in digital preservation. Journal of the Association for Information Science and Technology, 71(4):474–484, 2020.
2. M. Barons, S. Bhatia, J. Double, T. Fonseca, A. Green, S. Krol, H. Merwood, A. Mulinder, S. Ranade, J.Q. Smith, T. Thornhill, and D.H. Underdown. Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk. Archives and Records, 42(1):58–78, 2021.
3. Simon French. From soft to hard elicitation. *Journal of the Operational Research Society*, pages 1–17, 2021.
4. F. Jensen and T.D. Nielsen. Bayesian networks and decision graphs. Springer, 2007.
5. J.Q. Smith. Bayesian decision analysis: principles and practice. Cambridge University Press, 2010.
6. J.Q. Smith, M.J. Barons, and M. Leonelli. Coherent frameworks for statistical inference serving integrating decision support systems. arXiv preprint arXiv:1507.07394, 2015.
7. M.J. Barons, S.K. Wright, and J.Q. Smith. Eliciting probabilistic judgements for integrating decision support systems. Springer, New York, New York, USA, 2018.
8. M. Scutari and J.-B. Denis. Bayesian Networks: With Examples in R. CRC Press, 2014.
9. F. Bolger, A. Hanea, A. O'Hagan, O. Mosbach-Schulz, J. Oakley, G. Rowe, and M. Wenholt. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal, 12(6):Parma, Italy, 2014.
10. A. Hanea, M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, S. Mascaro, and B. Manning. $\mathrm{I}_{\mathrm{nvestigate}}\mathrm{D}_{\mathrm{iscuss}}\mathrm{E}_{\mathrm{stimate}}\mathrm{A}_{\mathrm{ggregate}}$ for structured expert judgement. International Journal of Forecasting, 33(1):267–279, 2016.
11. Roger Cooke, Max Mendel, and Wim Thijs. Calibration and information in expert resolution; a classical approach. Automatica, 24:87–93, 1988.

Computational Methods for Market Making Algorithms

Olivier Guéant

Abstract With the rise of electronification and trading automation, the task of quoting assets on many financial markets must be carried out algorithmically by market makers. Market making models and algorithms have therefore been an important research topic in recent years, at the frontier between economics, quantitative finance, scientific computing, and machine learning. The goal of this text is (i) to present a typical multi-asset market making model relevant for most over-the-counter markets, (ii) to show how to use stochastic optimal control tools to derive a theoretical characterization of optimal quotes in that model, and (iii) to discuss the various methods proposed in the literature that could be used in practice in the financial industry for building market making algorithms.

1 Introduction

In finance, securities may be traded through exchanges or over-the-counter (OTC) directly between two parties. In OTC markets, some market participants are systematically providing liquidity to the others by showing/answering prices at which they agree to buy and sell the assets and contracts they cover. These market participants are called dealers or market makers and they play a central role in the functioning of markets.

With the rise of electronification and trading automation, the task of quoting assets on OTC markets must be carried out algorithmically by market makers. Market making models and algorithms have therefore been an important research topic in recent years, at the frontier between economics, quantitative finance, scientific computing, and machine learning.

O. Guéant (✉)
Université Paris 1 Panthéon Sorbonne, Centre d'Economie de la Sorbonne, Paris, France
e-mail: olivier.gueant@univ-paris1.fr

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_66

In the 1980s, long time before algorithmic trading became necessary, economists proposed models where one or several risk-averse market makers optimized their pricing policy for managing their inventory risk (see [16–18]). More than twenty years later, Avellaneda and Stoikov revisited in [1] that literature with a quantitative finance viewpoint and proposed a model based on stochastic optimal control tools to help market makers determine their optimal quotes. This model inspired both academics and practitioners who then developed several realistic extensions—in particular for OTC markets (foreign exchange, bonds, etc.) although the paper was initially developed for stock markets. In [14], Guéant, Lehalle, and Fernandez-Tapia proved the first set of mathematical results on the Avellaneda-Stoikov model, derived closed-form approximations of the optimal quotes, and proposed extensions to include a drift in the price dynamics and adverse selection. Cartea and Jaimungal, along with several researchers added many features to the initial models: alpha signals, ambiguity aversion, etc. (see [7–10]). They also proposed a slightly different optimization framework where market makers maximize their expected profit minus a running penalty to avoid holding large inventories whereas [1] relied on an exponential utility function. To cite a few other extensions, general intensities and partial information in [6], persistence of the order flow in [19], multiple requested sizes in [4], client tiering and access to a liquidity pool in [2].

In most practical cases, market making algorithms must be built for entire portfolios whereas most models proposed in the literature have been single-asset ones until recently. Guéant and Lehalle were the first to touch upon a multi-asset extension of models *à la* Avellaneda-Stoikov in [13] and a complete analysis for the various objective functions present in the literature have been carried out in [12] (see also the book [11]). In spite of equations characterizing the optimal quotes, approximating numerically the optimal quotes remains a research problem, because of the curse of dimensionality. The goal of this chapter is to present a typical multi-asset market making model (Sect. 2), a theoretical characterization of the optimal quotes in that model (Sect. 3), and to discuss the various methods proposed in the literature that could be used in the financial industry (Sect. 4).

2 A Multi-Asset Market Making Model

We present here a typical model for the market making of $d \geq 1$ assets.

For $i \in \{1, \dots, d\}$, the reference price of asset i is modeled by a process $(S_t^i)_{t\in\mathbb{R}_+}$ with dynamics $dS_t^i = \sigma^i dW_t^i$ and S_0^i given, where $(W_t^1)_{t\in\mathbb{R}_+}, \dots, (W_t^d)_{t\in\mathbb{R}_+}$ are d Brownian motions with correlation matrix $(\rho^{i,j})_{1\leq i,j\leq d}$—hereafter we write $\Sigma = (\rho^{i,j}\sigma^i\sigma^j)_{1\leq i,j\leq d}$.

At each point in time, the market maker chooses the price at which she is ready to buy/sell each asset: for $i \in \{1, \dots, d\}$, we let her bid and ask quotes for asset i be modeled by two stochastic processes, respectively denoted by $(S_t^{i,b})_{t\in\mathbb{R}_+}$ and $(S_t^{i,a})_{t\in\mathbb{R}_+}$. For $i \in \{1, \dots, d\}$, we denote by $(N_t^{i,b})_{t\in\mathbb{R}_+}$ and $(N_t^{i,a})_{t\in\mathbb{R}_+}$ the two point processes modeling the number of transactions at the bid and at the ask, respectively, for asset i. In this simple model, the transaction size for asset i is constant and denoted by z^i. The inventory process of the market maker for asset i, denoted by $(q^i_t)_{t\in\mathbb{R}_+}$, has therefore the dynamics $dq^i_t = z^i dN^{i,b}_t - z^i dN^{i,a}_t$ with q^i_0 given, and we denote by $(q_t)_{t\in\mathbb{R}_+}$ the (column) vector process $(q^1_t, \dots, q^d_t)^\top_{t\in\mathbb{R}_+}$.

For each $i \in \{1, \dots, d\}$, we denote by $(\lambda^{i,b}_t)_{t\in\mathbb{R}_+}$ and $(\lambda^{i,a}_t)_{t\in\mathbb{R}_+}$ the intensity processes of $(N^{i,b}_t)_{t\in\mathbb{R}_+}$ and $(N^{i,a}_t)_{t\in\mathbb{R}_+}$, respectively.[1] We assume that the market maker stops proposing a bid (respectively ask) price for asset i when her position in asset i following the transaction would exceed a given threshold Q^i (respectively $-Q^i$). We assume that the intensities verify $\lambda^{i,b}_t = \Lambda^{i,b}(\delta^{i,b}_t)1_{\{q^i_{t-}+z^i\le Q^i\}}$ and $\lambda^{i,a}_t = \Lambda^{i,a}(\delta^{i,a}_t)1_{\{q^i_{t-}-z^i\ge -Q^i\}}$ where the processes $(\delta^{i,b}_t)_{t\in\mathbb{R}_+}$ and $(\delta^{i,a}_t)_{t\in\mathbb{R}_+}$ are defined by $\delta^{i,b}_t = S^i_t - S^{i,b}_t$ and $\delta^{i,a}_t = S^{i,a}_t - S^i_t$, for all $t \in \mathbb{R}_+$. Moreover, we assume that the functions $\Lambda^{i,b}$ and $\Lambda^{i,a}$ are twice continuously differentiable, decreasing[2] with $\forall\delta \in \mathbb{R}$, $\Lambda^{i,b/a\prime}(\delta) < 0$, and such that $\lim_{\delta\to+\infty}\Lambda^{i,b/a}(\delta) = 0$ and $\sup_\delta \frac{\Lambda^{i,b/a}(\delta)\Lambda^{i,b/a\prime\prime}(\delta)}{\left(\Lambda^{i,b/a\prime}(\delta)\right)^2} < 2$.

Finally, the process $(X_t)_{t\in\mathbb{R}_+}$ modelling the amount of cash on the market maker's cash account has the following dynamics:

$$dX_t = \sum_{i=1}^{d} S^{i,a}_t z^i dN^{i,a}_t - S^{i,b}_t z^i dN^{i,b}_t = \sum_{i=1}^{d}\left(\delta^{i,b}_t z^i dN^{i,b}_t + \delta^{i,a}_t z^i dN^{i,a}_t\right) - \sum_{i=1}^{d} S^i_t dq^i_t.$$

For the market maker, a classical optimization problem consists in maximizing the expected value of an exponential utility function (with risk aversion parameter $\gamma > 0$) applied to the mark-to-market (MtM) value of the portfolio at a given time T, i.e. the amount X_T plus the MtM value $\sum_{i=1}^d q^i_T S^i_T$ of the assets at time T:

$$\sup_{(\delta^{1,b}_t)_t,\dots,(\delta^{d,b}_t)_t,(\delta^{1,a}_t)_t,\dots,(\delta^{d,a}_t)_t\in\mathcal{A}} \mathbb{E}\left[-\exp\left(-\gamma\left(X_T + \sum_{i=1}^{d} q^i_T S^i_T\right)\right)\right],$$

where $\mathcal{A}$ is the set of predictable processes bounded from below. Alternatively, we can consider a risk-adjusted expectation for the objective function:

$$\sup_{(\delta^{1,b}_t)_t,\dots,(\delta^{d,b}_t)_t,(\delta^{1,a}_t)_t,\dots,(\delta^{d,a}_t)_t\in\mathcal{A}} \mathbb{E}\left[X_T + \sum_{i=1}^{d} q^i_T S^i_T - \frac{1}{2}\gamma\int_0^T q_t^\top \Sigma q_t\, dt\right].$$

Results for one of the two optimization problems usually translate into results for the other.

[1] Intensities are instantaneous probabilities to trade in this context.

[2] The probability to trade with a client depends monotonically on the proposed price.

3 Theoretical Results

The above problem is a stochastic optimal control problem that can be solved by using a Hamilton-Jacobi-Bellman (HJB) equation and a verification argument. In our case, the HJB equation is

$$
\begin{aligned}
0 = {} & \partial_t u(t,x,q,S) + \frac{1}{2}\sum_{i,j=1}^{d} \rho^{i,j}\sigma^i\sigma^j \partial^2_{S^i S^j} u(t,x,q,S) \\
& + \sum_{i=1}^{d} 1_{\{q^i+z^i \le Q^i\}} \sup_{\delta^{i,b}} \Lambda^{i,b}(\delta^{i,b}) \Big(u(t, x - z^i S^i + z^i \delta^{i,b}, q + z^i e^i, S) - u(t,x,q,S) \Big) \\
& + \sum_{i=1}^{d} 1_{\{q^i-z^i \ge -Q^i\}} \sup_{\delta^{i,a}} \Lambda^{i,a}(\delta^{i,a}) \Big(u(t, x + z^i S^i + z^i \delta^{i,a}, q - z^i e^i, S) - u(t,x,q,S) \Big),
\end{aligned}
$$

for all $(t,x,q,S) \in [0,T) \times \mathbb{R} \times \prod_{i=1}^{d} \left(z^i\mathbb{Z} \cap [-Q^i, Q^i] \right) \times \mathbb{R}^d$, where $\{e^i\}_{i=1}^{d}$ is the canonical basis of $\mathbb{R}^d$ and the terminal condition is

$$
u(T,x,q,S) = -\exp\left(-\gamma\left(x + \sum_{i=1}^{d} q^i S^i\right)\right), \forall (x,q,S) \in \mathbb{R} \times \prod_{i=1}^{d} \left(z^i\mathbb{Z} \cap [-Q^i, Q^i] \right) \times \mathbb{R}^d.
$$

Using the ansatz $u(t,x,q,S) = -\exp\left(-\gamma\left(x + \sum_{i=1}^{d} q^i S^i + \theta(t,q)\right)\right)$, it is straightforward to verify that solving the above HJB equation boils down to solving the following system of nonlinear ordinary differential equations:[3]

$$
\begin{aligned}
0 = {} & \partial_t \theta(t,q) - \frac{1}{2}\gamma q^\top \Sigma q \\
& + \sum_{i=1}^{d} 1_{\{q^i+z^i \le Q^i\}} z^i H_\xi^{i,b}\left(\frac{\theta(t,q) - \theta(t, q + z^i e^i)}{z^i} \right) \\
& + \sum_{i=1}^{d} 1_{\{q^i-z^i \ge -Q^i\}} z^i H_\xi^{i,a}\left(\frac{\theta(t,q) - \theta(t, q - z^i e^i)}{z^i} \right)
\end{aligned}
$$

with terminal condition $\theta(t,q) = 0$, where, for each $i \in \{1,\ldots,d\}$, $H^{i,b/a}(p) = \sup_\delta \frac{\Lambda^{i,b/a}(\delta)}{\gamma z^i}(1 - \exp(-\gamma z^i(\delta - p)))$.

Then, one can prove the following theorem using a verification argument (see [12], with slightly different notations):

[3] It is indeed a system of nonlinear ordinary differential equations because the variable q takes discrete values.

Theorem 1 *There exists a unique function* $\theta\colon [0,T]\times\prod_{i=1}^{d}(z^i\mathbb{Z}\cap[-Q^i,Q^i])\to\mathbb{R}$, C^1 *in time, solution of the above equation. Moreover, for* $i\in\{1,\ldots,d\}$, *the optimal bid and ask quotes are characterized by*

$$\delta_t^{i,b*}=\tilde{\delta}^{i,b*}\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}+z^ie^i)}{z^i}\right)\quad \textit{for}\ q_{t-}+z^ie^i\in\prod_{j=1}^{d}\left(z^j\mathbb{Z}\cap[-Q^j,Q^j]\right),$$

$$\delta_t^{i,a*}=\tilde{\delta}^{i,a*}\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}-z^ie^i)}{z^i}\right)\quad \textit{for}\ q_{t-}-z^ie^i\in\prod_{j=1}^{d}\left(z^j\mathbb{Z}\cap[-Q^j,Q^j]\right),$$

where the functions $\tilde{\delta}^{i,b*}(\cdot)$ *and* $\tilde{\delta}_{\gamma}^{i,a*}(\cdot)$ *are defined by*

$$\tilde{\delta}^{i,b/a*}(p)={\Lambda^{i,b/a}}^{-1}\left(\gamma z^iH^{i,b/a}(p)-H^{i,b/a'}(p)\right).$$

4 Numerical Methods

The above theorem states that finding the optimal quotes boils down to solving two problems:[4] (i) finding a numerical approximation of the function θ and (ii) computing the functions $\tilde{\delta}^{i,b/a*}$ for $i\in\{1,\ldots,d\}$. The latter problem does not raise any issue as the functions can be computed asset by asset by using classical optimization techniques. For the former, one needs to approximate numerically the solution of a system of nonlinear ordinary differential equations. For that purpose, two families of methods exist: grid methods where the solution is approximated at specific points and formula methods where the solution is approximated using simple or complex "combinations" of simple functions.

In the literature, it is common to see finite different methods on a grid to approximate θ. More precisely, Euler monotone schemes – explicit or implicit – are often used to solve this type of problems (see for instance [2, 4, 12]). Grid methods are very efficient in the one-asset case ($d=1$) or when d is small (say $d\leq 3$). However, because they require a grid of dimension $d+1$ (one dimension of time and d dimensions for the assets) grid methods naturally suffer from the curse of dimensionality and cannot be used for larger d.

Grid methods can nevertheless be used if one reduces beforehand the dimensionality of the problem. An interesting way proposed in [4] consists of (i) approximating the covariance matrix Σ by a low-rank symmetric matrix by using a principal component analysis and keeping $k\leq 3$ risk factors—and therefore replacing the d-dimensional variable q by a low-dimensional one corresponding to the k risk factors—and (ii) replacing the risk limits in terms of assets $(Q^i)_i$ by risk

[4] For most extensions of the above model, these two problems remain the relevant ones.

limits in terms of factor exposures. By using this approximation, θ can be regarded as a function of time and risk factors and approximated using a grid of dimension $k + 1$ and not $d + 1$.

To beat the curse of dimensionality, formula methods can of course be used. Closed-form formulas have been proposed in [12] and more recently in [5]. In [5], the idea was to "approximate" the system of nonlinear ordinary differential equations by a multi-dimensional Riccati equation that can be solved in closed form. The approximation of θ turns out to be a polynomial of degree 2 in that case and approximations of the optimal quotes are then derived asset by asset using the above equations. This type of techniques provides great results, as exemplified in [5].

In order to approximate the optimal quotes using a formula method, another interesting idea consists in looking for the function θ and sometimes the optimal quotes themselves in the form of neural networks using reinforcement learning techniques. Promising results in this line can be found in [15].

Numerical examples regarding credit indices are presented in [12]. For bonds, the papers [4] and [15] contain interesting illustrations. The case of foreign exchange has been tackled recently in [3] and should attract more interest in the near future.

References

1. M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. Quantitative Finance, 8(3):217–224, 2008.
2. A. Barzykin, P. Bergault, and O. Guéant. Algorithmic market making in foreign exchange cash markets with hedging and market impact. Working paper, 2021.
3. A. Barzykin, P. Bergault, and O. Guéant. Market making by an FX dealer: tiers, pricing ladders and hedging rates for optimal risk control. Working paper, 2022.
4. P. Bergault and O. Guéant. Size matters for OTC market makers: general results and dimensionality reduction technique. Mathematical Finance, 31(1):279–322 2020.
5. P. Bergault, D. Evangelista, O. Guéant, and D. Vieira. Closed-form approximations in multi-asset market making. Applied Mathematical Finance, to appear.
6. L. Campi and D. Zabaljauregui. Optimal market making under partial information with general intensities. Applied Mathematical Finance 27, 2020.
7. Á.Cartea, R. Donnelly, and S. Jaimungal. Algorithmic trading with model uncertainty. SIAM Journal on Financial Mathematics, 8(1):635–671, 2017.
8. Á. Cartea, S. Jaimungal, and J. Penalva. Algorithmic and high-frequency trading. Cambridge University Press, 2015.
9. Á. Cartea, S. Jaimungal, and J. Ricci. Buy low, sell high: A high frequency trading perspective. SIAM Journal on Financial Mathematics, 5(1):415–444, 2014.
10. Á. Cartea, S. Jaimungal, and J. Ricci. Algorithmic trading, stochastic control, and mutually exciting processes. *SIAM Review*, 60(3):673–703, 2018.
11. O. Guéant. The Financial Mathematics of Market Liquidity: From optimal execution to market making, volume 33. CRC Press, 2016.
12. O. Guéant. Optimal market making. Applied Mathematical Finance, 24(2):112–154, 2017.
13. O. Guéant and C.-A. Lehalle. General intensity shapes in optimal liquidation. Mathematical Finance, 25(3):457–495, 2015.
14. O. Guéant, C.-A. Lehalle, and J. Fernandez-Tapia. Dealing with the inventory risk: a solution to the market making problem. Mathematics and Financial Economics, 7(4):477–507, 2013.

15. O. Guéant and I. Manziuk. Deep reinforcement learning for market making in corporate bonds: beating the curse of dimensionality. Applied Mathematical Finance, 26(5):387–452, 2019.
16. Th. Ho and H.R. Stoll. On dealer markets under competition. The Journal of Finance, 35(2):259–267, 1980.
17. Th. Ho and H.R. Stoll. Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics, 9(1):47–73, 1981.
18. Th. Ho and H.R. Stoll. The dynamics of dealer markets under competition. The Journal of Finance, 38(4):1053–1074, 1983.
19. P. Jusselin. Optimal market making with persistent order flow. arXiv preprint arXiv:2003.05958, 2020.

Study of Self-Adjoint Singularly Perturbed BVP by Septic Hermite Collocation Method

Archna Kumari, Shallu Shallu, and V. K. Kukreja

Abstract Self-adjoint singular perturbed boundary value problems are analysed using orthogonal collocation on finite elements with septic Hermite as a basis function. The roots of shifted Legendre polynomials are taken as collocation points. After discretization using septic Hermite collocation method, the BVP reduces to a banded system of $6N \times 6N$ linear equations. The proposed method is found to be stable and has an order of convergence of six. To validate the accuracy of the method, SHCM is applied on few test problems. The method exhibits excellent results for very small value of the perturbation parameter to the range of 2^{-1000} and beyond. This demonstrates the efficiency and reliability of the method.

1 Introduction

Consider the following self-adjoint singularly perturbed boundary value problem (BVP),

$$\mathbf{L}_\varepsilon \equiv \varepsilon y'' + m(x)y' + n(x)y = p(x), \quad x \in (a, b), \tag{1}$$

with condition:

$$y(a) = A, \quad y(b) = B, \tag{2}$$

where $0 < \varepsilon << 1$ is a small perturbation parameter and $m(x)$, $n(x)$ and $p(x)$ are sufficiently smooth functions, such that $m(x) > m^* > 0$ and $n(x) > n^* > 0$. Under these conditions Eqs. (1)–(2) have a unique solution. In general, the solution $y(x)$ may exhibit two boundary layers of exponential type at both end points $x = a$ and $x = b$.

A. Kumari · S. Shallu · V. K. Kukreja (✉)
Department of Mathematics, SLIET Longowal, Longowal, Punjab, India

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_67

Self-adjoint singular perturbation problems arises in the various fields of science and engineering such as mass and heat transfer problems [1], fluid dynamics [2], chemical and biological dynamics [3], elasticity, and optimal control [4]. It is a well-known fact that the solution of self-adjoint singularly perturbed boundary-value problem displays a multiscale character. In some regions, i.e., a thin layer where the solution varies rapidly, while behaves regularly and slowly when away from the layer solution. When the perturbation parameter $\varepsilon \to 0$, major computational difficulties arise and the standard methods do not yield accurate results for all value of x. Many numerical methods are available in the literature to solve the second-order self-adjoint singular perturbation problems [5, 7]. For details, one can refer to the survey article by Kadalbajoo and Patidar [8].

2 Description of Septic Hermite Collocation Method

Septic Hermite collocation method (SHCM) is one of the weighted residual methods. It is a combination of the finite element method and orthogonal collocation method with septic Hermite as the basis function, which is C^3 continuous. Let $\pi = \{a = x_0 < x_1 < \cdots < x_N = b\}$ be the partition of given domain $[a, b]$ into N number of subdomain called finite element, with uniform step size $h = (b - a)/N$. Now, the physical solution $y(x)$ over the domin $[a, b]$ can be approximated by a piecewise Hermite polynomials of 7th degree as follows:

$$\bar{y}(x) = \sum_{k=0}^{N} [a_k A_k(x) + b_k B_k(x) + c_k C_k(x) + d_k D_k(x)],$$

where a_k, b_k, c_k, d_k are unknown constants and $A_k(x)$, $B_k(x)$, $C_k(x)$ and $D_k(x)$ are septic Hermite interpolating polynomials of degree seven. The details of these polynomials are given in [9].

In SHCM, the given domain $[a, b]$ is discretized into N number of sub-domains called finite element and h is uniform spacing. After the discretization, each element $[x_k, x_{k+1}]$ is mapped to $[0, 1]$ by using the transformation $\eta = \frac{x - x_k}{h}$. Further, roots of sixth degree shifted Legendre polynomials are used as collocation points. Six interior collocation points within each element $[x_k, x_{k+1}]$ are introduced. More details about SHCM is given in [9].

2.1 Choice of Trial Function and Discretization Process

The approximate solution $\bar{y}$ of exact solution y in k^{th} element can be written as:

$$\bar{y}(x) = \sum_{i=1}^{8} a_{i+6k-6} H_i(\eta), \quad 0 \le \eta \le 1, \tag{3}$$

where $a_i's$ are the unknown variables and $k = 1, 2, \dots, N$. Using the approximate solution (3) in the Eqs. (1)–(2) at q^{th} collocation point, the discretized set of equations can be written as:

$$\sum_{i=1}^{8} a_{i+6k-6}\Big[\varepsilon H_i''(\eta_q)+m(\eta_q h+x_k)hH_i'(\eta_q)+n(\eta_q h+x_k)h^2H_i(\eta_q)\Big] = h^2p(\eta_q h+x_k), \tag{4}$$

where $q = 1, 2, \dots, 6$. The system (4) consist of $6N$ equations involving $6N + 2$ unknowns. Two extra unknowns are calculated using boundary conditions $y(a) = A$ and $y(b) = B$. After using the boundary conditions, the system (4) reduces to matrix form as:

$$\mathbb{S}\mathbb{X}^N = \mathbb{L}^N, \tag{5}$$

where $\mathbb{S}$ is coefficient matrix and $\mathbb{X}^N = [a_2, a_3, \dots, a_{6N}, a_{6N+2}]^T$ are the unknowns vector to be determined and $\mathbb{L}^N$ is the column vector. The matrix $\mathbb{S}$ is diagonally dominant and non-singular. The matrix system (5) is solved using MATLAB software. However, the matrices $\mathbb{S}$ and $\mathbb{L}^N$ for $i = 1, 2, \dots, 8$, $q = 1, 2, \dots, 6$ are defined as follows:

$$\mathbb{S} = \begin{cases} \varepsilon H_i''(\eta_q) + hm(\eta_q h + x_1)H_i'(\eta_q) + n(\eta_q h + x_1)h^2H_i(\eta_q), & \text{at } k = 1,\ i \neq 1 \\ \varepsilon H_i''(\eta_q) + hm(\eta_q h + x_k)H_i'(\eta_q) + n(\eta_q h + x_k)h^2H_i(\eta_q), & \text{at } k = 2, 3, \dots, N-1, \\ \varepsilon H_i''(\eta_q) + hm(\eta_q h + x_N)H_i'(\eta_q) + n(\eta_q h + x_N)h^2H_i(\eta_q), & \text{at } k = N, i \neq 7 \end{cases}$$

$$\mathbb{L}^N = \begin{cases} h^2p(\eta_q h + x_1) - A(\varepsilon H_1''(\eta_q) + hm(\eta_q h + x_1)H_1'(\eta_q) & \\ \quad +n(\eta_q h + x_1)h^2H_1(\eta_q)), & \text{at } k = 1, \\ h^2p(\eta_q h + x_k), & \text{at } k = 2, 3, \dots, N-1, \\ h^2p(\eta_q h + x_N) - B(\varepsilon H_7''(\eta_q) + hm(\eta_q h + x_N)H_7'(\eta_q) & \text{at } k = N \\ \quad +n(\eta_q h + x_N)h^2H_7(\eta_q)), & \end{cases}$$

3 Stability Analysis

Let $\theta\mathbb{S}$, $\theta\mathbb{L}^{\mathrm{N}}$ is the inbuilt error in the calculation of $\mathbb{S}$ and $\mathbb{L}^N$ respectively and suppose $\mathcal{X}^N$ be the solution of the system (5), i.e.,

$$(\mathbb{S} + \theta\mathbb{S})\mathcal{X}^N = \mathbb{L}^N + \theta\mathbb{L}^N. \tag{6}$$

Septic Hermite collocation method is said to be stable if ∃ non-negative constants L_1, L_2, L_3 such that the system (5) has a unique solution for $||\theta\mathbb{S}|| \leq L_3$ and

$$||\mathbb{X}^N - \mathcal{X}^N|| \leq (L_1||\theta\mathbb{S}||\,||\mathbb{X}^N|| + L_2||\theta\mathbb{L}^N||). \tag{7}$$

Since the matrix $\mathbb{S}$ is diagonally dominant, therefore, using the result of [11]:

$$||\mathbb{S}||^{-1} \leq \frac{1}{min\{7(\varepsilon/h^2 + m^*/h + n^*), 8(\varepsilon/h^2 + m^*/h + n^*)\}} \leq \frac{L}{h^2} = v, \tag{8}$$

where $m_i \geq m^* > 0$, $n_i \geq n^* > 0$ and maximum value of all septic Hermite basis function is less than or equal to one. Thus

$$||\mathbb{X}^N|| \leq ||\mathbb{S}^{-1}||\,||\mathbb{L}^N|| \leq L. \tag{9}$$

Choose a positive constant $u < (1/2)v$, then whenever $||\theta\mathbb{S}|| \leq u$, Eq. (6) has unique solution for

$$||(\mathbb{S} + \theta\mathbb{S})^{-1}|| = ||(I + \mathbb{S}^{-1}\theta\mathbb{S})^{-1}\mathbb{S}^{-1}|| \leq 2v,$$

because $||\mathbb{S}^{-1}\theta\mathbb{S}|| \leq ||\mathbb{S}^{-1}||\,||\theta\mathbb{S}|| \leq \frac{1}{2}$. Since $(\mathbb{S}+\theta\mathbb{S})(\mathbb{X}^N - \mathcal{X}^N) = \theta\mathbb{S}\mathbb{X}^N - \theta\mathbb{L}^N$, therefore,

$$\begin{aligned} ||\mathbb{X}^N - \mathcal{X}^N|| &\leq ||\mathbb{S} + \theta\mathbb{S}||^{-1}(||\theta\mathbb{S}\mathbb{X}^N - \theta\mathbb{L}^N||) \\ &\leq 2v(||\theta\mathbb{S}||\,||\mathbb{X}^N|| + ||\theta\mathbb{L}^N||), \end{aligned}$$

which ensures the stability of the septic Hermite collocation system.

4 Convergence Analysis

Theorem 1 ([12]) *Let $y(x)$ be the solution of Eqs. (1)–(2) such that $y(a) \geq 0$ and $y(b) \geq 0$. Then $\mathbf{L}_\varepsilon y \geq 0$, $\forall x \in (a, b)$ implies that $y(x) \geq 0$, $\forall\, x \in [a, b]$.*

Theorem 2 ([12]) *Let $y(x)$ be the solution of Eqs. (1)–(2). Then,*

$$\|y(x)\| \leq K_\varepsilon\big(\frac{\|p\|}{n_0} + max(|A|, |B|)\big), \quad \forall x \in [a, b],$$

where $0 < n_0 < n(x)$ $\forall\, x \in (a, b)$ and $\|.\|$ is maximum norm.

Theorem 3 ([12]) *Let $y(x)$ be the solution of Eqs. (1)–(2) and $m(x)$, $n(x)$ and $p(x)$ are sufficiently smooth functions in $[a, b]$, then $\exists$ a constant K_ε such that:*

$$|y^{(j)}| \leq K_\varepsilon, \quad \forall x \in [-1, -\omega)\bigcup(-\omega, 1], \quad j = 0, 1, \ldots, 4.$$

The bounds on the solution and its derivatives in the layer region $[-\omega, 0) \bigcup [0, \omega]$ are provided in the next two theorems.

Theorem 4 ([12]) *Let $y(x)$ be the solution of Eqs. (1)–(2) and $p(x)$ sufficiently smooth function in $[a, b]$, then $\exists$ a constants K_ε and $\varpi > 0$ such that:*

$$|y^{(j)}(x)| \leq K_\varepsilon \Big[1 + \varepsilon^{-j} e^{\varpi x/\varepsilon}\Big], \quad \forall x \in [-\omega, 0), \quad j = 1, 2, \ldots$$

Theorem 5 ([12]) *Let $y(x)$ be the solution of Eqs. (1)–(2) and $p(x)$ sufficiently smooth function in $[a, b]$, then $\exists$ a constants K_ε and $\varpi > 0$ such that:*

$$|y^{(j)}(x)| \leq K_\varepsilon \Big[1 + \varepsilon^{-j} e^{-\varpi x/\varepsilon}\Big], \quad \forall x \in [0, \omega], \quad j = 1, 2, \ldots$$

Theorem 6 ([10]) *Let $\bar{y}(x)$ be the septic Hermite splines approximation from the space $\breve{\mathbb{H}}$ to the solution $y(x)$ of the Eqs. (1)–(2). If $p(x) \in C^2[0, 1]$, then the uniform error estimate is given by:*

$$\| y - \bar{y} \|_\infty \leq K_\varepsilon h^6.$$

5 Results and Discussion

The maximum absolute error of each example are calculated by using the formula:

$$E_\varepsilon^N = \max_{1 \leq i \leq N+1} |y^{\text{ex}}(x_i) - y^{\text{app}}(x_i)|$$

where $y^{\text{ex}}(x_i)$ denotes the exact solution and $y^{\text{app}}(x_i)$ denotes the numerical solution of given Eqs. (1)–(2).

Example 1 Consider the self-adjoint singularly perturbed problem [6, 7] with the boundary conditions:

$$-\varepsilon y''(x) + y(x) = -(\cos^2(\pi x) + 2\varepsilon\pi^2 \cos(2\pi x)), \quad x \in [0, 1]$$
$$y(0) = 0, \quad y(1) = 0.$$

The exact solution of the given problem is:

$$y(x) = \frac{e^{\frac{-(1-x)}{\sqrt{\varepsilon}}} + e^{\frac{-x}{\sqrt{\varepsilon}}}}{1 + e^{\frac{-1}{\sqrt{\varepsilon}}}} - \cos^2(\pi x)$$

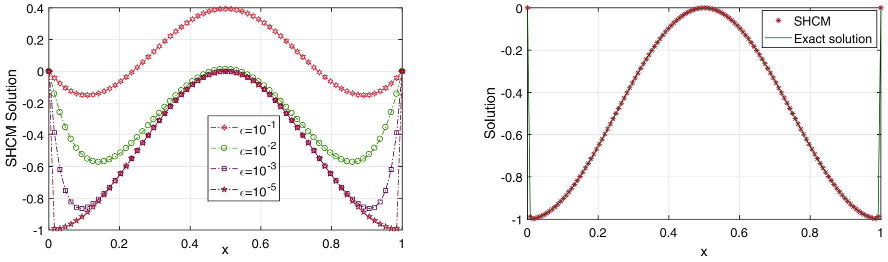
The estimated maximum absolute error is presented in Table 1 and the comparison of maximum absolute error for different values of ε and N with fitted finite

Table 1 Maximum absolute error of Example 1 for different values of ε and N

ε	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
2^{-15}	1.668843E−03	8.050322E−05	5.145507E−07	6.489208E−10	2.864375E−13	5.440092E−15
2^{-16}	3.830381E−03	4.718962E−04	8.171656E−06	2.156443E−08	1.504674E−11	6.439293E−15
2^{-17}	6.347774E−03	1.668843E−03	8.050322E−05	5.145507E−07	6.489208E−10	2.866595E−13
2^{-18}	8.456865E−03	3.830381E−03	4.718962E−04	8.171656E−06	2.156443E−08	1.504674E−11
2^{-19}	9.872092E−03	6.347774E−03	1.668843E−03	8.050322E−05	5.145507E−07	6.489205E−10
2^{-20}	1.070167E−02	8.456865E−03	3.830381E−03	4.718962E−04	8.171656E−06	2.156443E−08
2^{-25}	1.159884E−02	1.150761E−02	1.115237E−02	9.872092E−03	6.347774E−03	1.668843E−03
2^{-30}	1.162852E−02	1.162564E−02	1.161415E−02	1.156832E−02	1.138749E−02	1.070167E−02
2^{-50}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-100}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-500}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-1000}	1.162947E−02	1.162947E−02	1.162947E−02	1.1629479E−02	1.162947E−02	1.162947E−02

Table 2 Comparisons of maximum absolute error of Example 1 for different values of ε and N

	$N = 8$	$N = 16$	$N = 32$	$N = 64$
SHCM				
$\varepsilon = 10^{-3}$	1.922453E−08	1.317701E−11	5.551115E−15	7.105427E−15
$\varepsilon = 10^{-4}$	1.413032E−04	1.192632E−06	1.831640E−09	9.100498E−13
$\varepsilon = 10^{-5}$	5.379017E−03	1.078566E−03	3.501474E−05	1.558968E−07
Soujanya and Reddy[6]				
$\varepsilon = 10^{-3}$	6.15E−02	1.38−E02	3.28E−03	8.07E−04
$\varepsilon = 10^{-4}$	7.15E−02	1.72E−02	3.76E−03	8.47E−04
$\varepsilon = 10^{-5}$	7.30E−02	1.88E−02	4.60E−03	1.04E−03



(a) (b)

Fig. 1 Exact and approximate solution of Example 1 for (**a**) N = 64. (**b**) $\varepsilon = 10^{-10}$, N = 128

difference method [6] is given in Table 2. The graphical behavior of the numerical and exact solution is given in Fig. 1.

Example 2 Consider the self-adjoint singularly perturbed problem [6] with the boundary conditions:

$$-\varepsilon y''(x) + (2 - x^2)y(x) = 1, \quad x \in [-1, 1]$$
$$y(-1) = 0, \quad y(1) = 0.$$

The exact solution of the given problem is:

$$y(x) = \frac{1}{(2 - x^2)} - e^{\frac{-(1+x)}{\sqrt{\varepsilon}}} - e^{\frac{-(1-x)}{\sqrt{\varepsilon}}}.$$

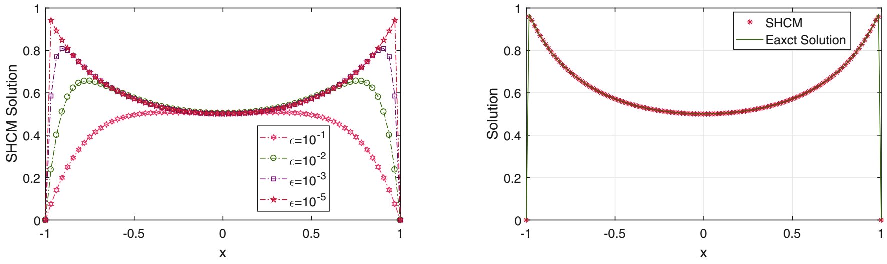
Table 3 shows the maximum absolute error for different values of ε and N. A comparison of maximum absolute error with fitted finite difference method [6] is reported in Table 4. Figure 2 shows the behavior of numerical and exact solutions for different ε and N.

Table 3 Maximum absolute error of Example 2 for different values of ε and N

ε	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
2^{-15}	2.551540E−03	2.570550E−04	4.258611E−06	1.399122E−08	1.461375E−11	7.105427E−15
2^{-16}	4.779032E−03	9.872312E−04	4.157101E−05	2.844487E−07	5.126749E−10	3.188560E−13
2^{-17}	7.149446E−03	2.537891E−03	2.545537E−04	4.200545E−06	1.377199E−08	1.436062E−11
2^{-18}	9.145310E−03	4.766763E−03	9.820070E−04	4.121909E−05	2.814405E−07	5.066375E−10
2^{-19}	1.049243E−02	7.141186E−03	2.531066E−03	2.533030E−04	4.171513E−06	1.366238E−08
2^{-20}	1.119786E−02	9.140990E−03	4.760629E−03	9.793949E−04	4.104313E−05	2.799364E−07
2^{-25}	1.162181E−02	1.159877E−02	1.147554E−02	1.048945E−02	7.133959E−03	2.525094E−03
2^{-30}	1.162923E−02	1.162852E−02	1.162564E−02	1.161414E−02	1.156576E−02	1.119696E−02
2^{-50}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-100}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-500}	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02	1.162947E−02
2^{-1000}	1.162947E−02	1.162947E−02	1.162947E−02	1.162948E−02	1.162947E−02	1.162947E−02

Table 4 Comparisons of maximum absolute error of Example 2 for different values of ε and N

	$N = 8$	$N = 16$	$N = 32$	$N = 64$
SHCM				
$\varepsilon = 10^{-3}$	1.989101E-03	8.425183E−03	1.644760E−02	1.644760E−02
$\varepsilon = 10^{-4}$	2.398622E−03	1.986540E−04	9.628221E−04	3.386290E−03
$\varepsilon = 10^{-5}$	9.570772E−03	5.578023E−03	1.100222E−03	5.874330E−05
Soujanya and Reddy[6]				
$\varepsilon = 10^{-3}$	1.89E−02	1.56E−02	1.65E−02	1.68E−02
$\varepsilon = 10^{-4}$	1.61E−02	6.09E−03	3.37E−03	3.78E−03
$\varepsilon = 10^{-5}$	1.61E−02	5.96E−03	1.87E−03	6.50E−04



(a)

(b)

Fig. 2 Exact and approximate solution of Example 2 for (**a**) N = 64. (**b**) $\varepsilon = 10^{-10}$, N = 128

6 Conclusion

In this work, the septic Hermite collocation method is developed to solve the self-adjoint singular perturbation problems. From the theoretical analysis, it can be found that the proposed method is stable and has sixth-order convergence. SHCM with uniform mesh is applied to two different problems. The numerical results exhibit that the proposed technique is efficient, reliable, and works even for a very small perturbation parameter (ε), which is never reported in the literature.

Acknowledgments Ms. Archna Kumari is thankful to MHRD New Delhi for providing financial assistance under TEQIP-III project. Ms. Shallu is thankful to CSIR New Delhi for providing financial assistance in the form of JRF file no. 09/797(0016)/2018-EMR-I.

References

1. Yaglom, A.M., Kader, B.A.: Heat and mass transfer between a rough wall and turbulent fluid flow at high Reynolds and Peclet numbers. J. Fluid Mech. **62**, 601-623 (1974).
2. Kan-On, Y., Mimura, M.: Singular perturbation approach to a 3-component reaction-diffusion system arising in population dynamics. SIAM J. Math. Anal. **29**, 1519-1536 (1988).

3. McGough, J.S., Riley, K.L.: A priori bounds for reaction-diffusion systems arising in chemical and biological dynamics. Appl. Math. Comput. **163**, 1-16 (2005).
4. Zhang, Y., Naidu, D.S., Cai, C., Zou, Y.: Singular perturbations and time scales in control theories and applications: an overview 2002–2012. Int. J. Inf. Syst. Sci. **9**, 1-36 (2014).
5. Ramadan M.A., Lashien I.F., Zahra W.K.: The numerical solution of singularly perturbed boundary value problems using nonpolynomila spline. Int. J. Pure Appl. Math. **6**, 883-896 (2007).
6. Soujanya G.B.S.L., Reddy Y.N.: Numerical solution of singular perturbation problems exhibiting dual layers. Int. J. Adv. Eng. Sci. Appl. Math. **5**, 250–257 (2013).
7. Lodhi R.K., Mishra H.K.: Septic B-spline method for second order self-adjoint singularly perturbed boundary-value problems. Ain Shams Eng. J. **9**, 2153–2161 (2018).
8. Kadalbajoo M.K., Patidar K.C.: A survey of numerical techniques for solving singularly-perturbed ordinary differential equations. Appl. Math. Comput. **13**, 457-510 (2002).
9. Kumari A., Kukreja V.K.: Robust septic Hermite collocation technique for singularly perturbed generalized Hodgkin–Huxley equation. Int. J. Comp. Math. **99**, 1–20 (2021).
10. Kumari A., Kukreja V.K.: Error bounds for septic Hermite interpolation and its implementation to study modified Burgers' equation. Num. Alg. **89**, 1799–1821 (2022).
11. Varah J.M.: A lower bound for the smallest singular value of a matrix. Linear Algebra Appl. **11**, 3–5 (1975).
12. Munyakazi J.B., Patidar K.C., Sayi M.T.: A robust fitted operator finite difference method for singularly perturbed problems whose solution has an interior layer. Math. Comput. Simul. **160**, 155–167 (2019).

Data-Driven Approach for Systemic Risk: A Macroprudential Perspective

Flavia Barsotti

Abstract This paper proposes a sovereign CDS analysis for systemic risk, assuming a macroprudential perspective and building on the modelling framework proposed by Baglioni and Cherubini (J. Econ. Dynam. Control 37:1581–1597, 2013). A data-driven approach applied to CDS quotes is considered to estimate a reduced form model for the marginal intensity of defaults at country level and investigate the presence of common factors. Results show a systematic effect on default intensities, rank correlation and common factors for countries in the sample with specific geographic differences. This is an important empirical evidence to further investigate how to model, measure and assess the drivers explaining heterogeneity in impacts across countries and build early warning indicators to support strategic decision making.

1 Introduction

From a macroprudential perspective, policy decision makers have faced challenging times all over the world to contain and manage the unprecedented effects caused by Covid-19 pandemic and build trust. Some countries have taken the decision of strict lockdown measures which in turns might have caused contraction in the economy for different sectors. As the economic fundamentals weaken, risk aversion begins to play a predominant role for agents [3]. The increased uncertainty affecting

The views expressed in this paper are solely those of the author and do not necessarily represent those of her current or past employers.

F. Barsotti (✉)
ING Analytics, Amsterdam, The Netherlands

IAS (Institute for Advanced Study), University of Amsterdam, Amsterdam, The Netherlands

DIAM (Delft Institute of Applied Mathematics), TU Delft, Delft, The Netherlands
e-mail: Flavia.Barsotti@ing.com; f.barsotti@uva.nl; F.Barsotti@tudelft.nl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_68

the world since 2019 is reflected in impacts on CDS spreads, as market-implied measure embedding default risk. This economic situation could potentially trigger cascading defaults for sovereigns. In order to build a preliminary empirical evidence at EU level, this paper entails an empirical study to investigate the dynamics of sovereign CDS and systemic risk from a macroprudential perspective, focusing on the credit risk component captured by CDS quotes and potential joint default effects. The paper proposes a data-driven approach to analyse the dynamics of sovereign CDS and systemic risk. Starting from the works by [1] and [6], a data-driven approach applied to CDS quotes is considered to estimate a reduced form model for the marginal intensity of defaults at country level. This information is then used to investigate the presence of common factors across economies for 8 European countries. Results show a systematic impact on pair-wise rank correlations and co-movements of the series for CDS sovereign market-implied indicators and marginal probabilities estimates. This is an important empirical evidence suggesting to build more extensive analysis to identify the factors explaining heterogeneity across countries by disentangling the default risk components and assessing the specific role of both systematic and non-systematic drivers [2].

2 Macroprudential Perspective: Risk Decomposition

The paper proposes a data-driven approach to investigate the implications on CDS market indicators in terms of credit quality and default risk of sovereigns close to the pandemic event outbreak. A key factor for macroprudential policy is building an holistic view on the financial system. This is fundamental to prevent overlooking the dependencies and impacts of its inner working mechanism and manage potential costs, instability and systemic risk patterns. From a mathematical point of view, this paper tackles the problem of estimating a reduced form approach for the drivers underlying the probability of a systemic risk event by leveraging on the marginal default probabilities of a set of obligors (e.g. sovereign), the pair-wise correlation of intensities and the identification of a common factor. The analysis of non-systematic components is beyond the scope of the present paper.

2.1 A Reduced Form Model for Sovereign Default Risk

As in [1], the paper considers a reduced form model where the default probability of each obligor i follows a Poisson process with intensity $\hat{\lambda}_i$. The intensity captures the instantaneous relative increase in the probability of an event, namely

$$dP_i(t) = \hat{\lambda}_i P_i(t), \tag{1}$$

with $P_i(t)$ being the probability of default at time t. The survival probability of a generic obligor i at time T is

$$P_i(\phi_i > T) = \exp\big(-\hat{\lambda}_i(T-t)\big), \tag{2}$$

with ϕ_i indicating the default time and $\{\hat{\lambda}_i, \hat{\lambda}_j\}$ the marginal sovereign default intensities of obligors $\{i, j\}$. Following the modelling dependence framework in [1, 6], we consider a *common factor* F and estimate its intensity $\bar{\lambda}_{F,ij}$ as

$$\bar{\lambda}_{F,ij} = \frac{2\hat{\rho}_{ij}(\hat{\lambda}_i + \hat{\lambda}_j)}{3(1-\hat{\rho}_{ij})}, \tag{3}$$

with $\hat{\rho}_{ij}$ being an estimate of the pair-wise correlation ρ_{ij}. For each rank correlation, we can then have an estimate of the common factor intensity[1] and analyze the impacts per geographic location.

2.2 Data

The analysis considers a sample of weekly observations of the most liquid 5Y CDS quotes for European sovereign CDS over the period Jan-2016/Jan-2021. Based on the financial nature of CDS contracts, their quotes enable to isolate the credit risk associated to its reference entity. In the case of sovereign, this plays a fundamental role for systemic risk measurement and assessment. In line with [6], this paper considers 8 European countries belonging to Northern Europe (e.g. Germany, France, Netherlands, UK—*Northern EU*) and Southern Europe (e.g. Greece, Italy, Portugal, Spain—*Southern EU*). As in [1], following the standard market approximation, the generic marginal intensity (e.g. hazard rate) $\hat{\lambda}_{C_{jt}}$ for sovereign C_j at time t is estimated as

$$\hat{\lambda}_{C_{jt}} = \frac{CDS_{jt}}{LGD}, \tag{4}$$

with CDS_{jt} denoting the CDS quote for obligor j at time t and LGD the loss given default. Figure 1 reports the time series of hazard rates for CDS sovereign obligors in the sample and Table 1 the associated descriptive statistics.

3 Empirical Evidence

The empirical evidence deriving from the estimation results highlights a certain degree of heterogeneity in the impacts of the pandemic outbreak in the sample.

[1] As estimate of the common factor F, this first empirical analysis considers a $30\%-$quantile measure over the set of estimates.

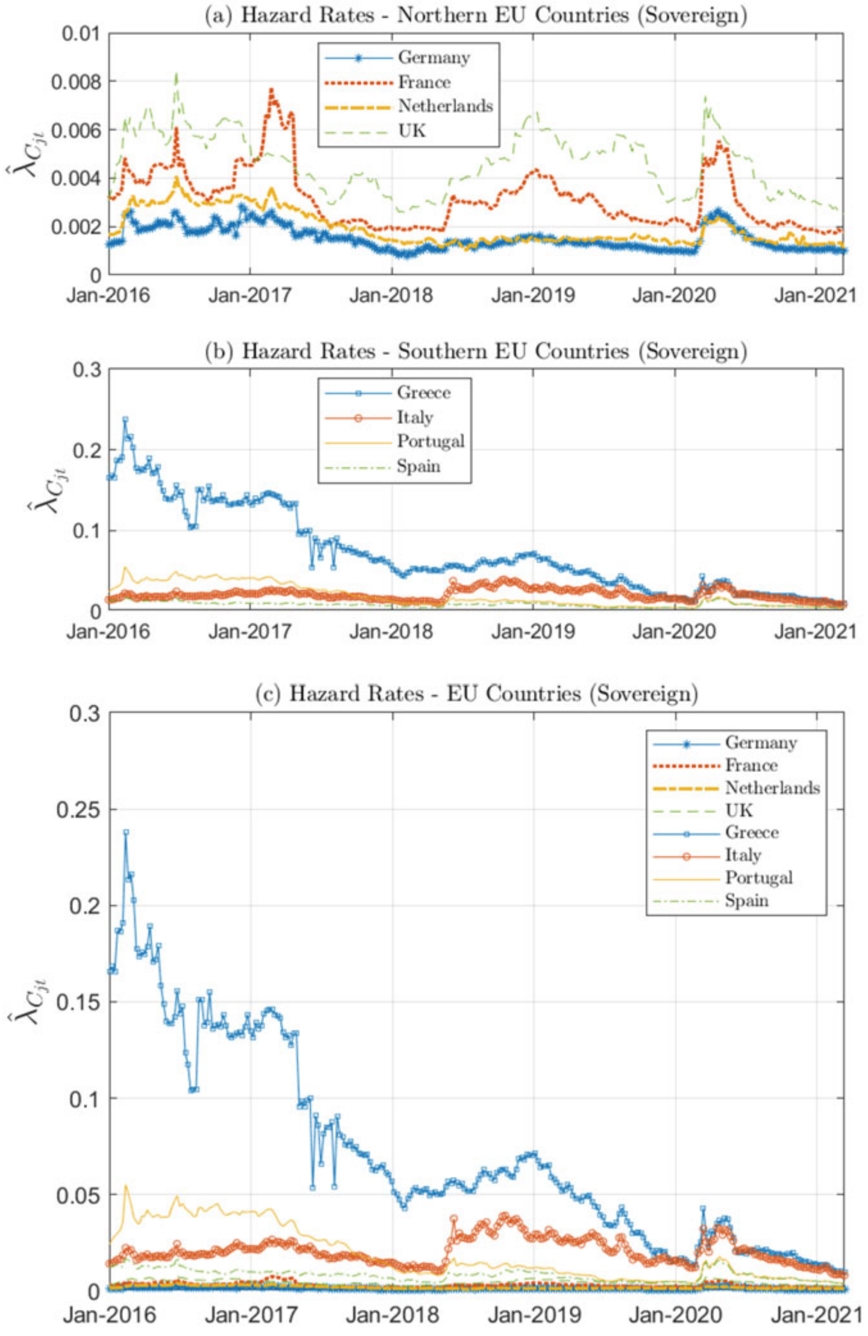


Fig. 1 Sovereign hazard rates. The plots report the hazard rates implied by CDS quotes for 8 EU Sovereign entities over the period Jan-2016/Jan-2021: Germany, France, Netherlands, UK, Greece, Italy, Portugal, Spain. The marginal intensity for Sovereign C_j at time t is measured via the hazard rate $\hat{\lambda}_{C_{jt}}$ defined in Eq. (4). Plots (**a**)–(**b**) report the hazard rates with a split by region, e.g. Northern EU Countries (**a**), Southern EU Countries (**b**). Plot (**c**) reports the comprehensive overview for all countries. Table 1 reports the corresponding descriptive statistics

Table 1 Descriptive statistics. Sovereign hazard rates implied by CDS quotes. Time window: Jan-2016/Jan-2021

Northern EU Sovereign	*Mean*	*Std dev*	*Min*	*Max*
Germany	0.0015	0.0005	0.0008	0.0028
France	0.0031	0.0012	0.0017	0.0078
Netherlands	0.0019	0.0007	0.0010	0.0041
UK	0.0046	0.0012	0.0025	0.0084
Southern EU Sovereign	*Mean*	*Std dev*	*Min*	*Max*
Greece	0.0717	0.0520	0.0098	0.2379
Italy	0.0207	0.0067	0.0082	0.0392
Portugal	0.0185	0.0139	0.0039	0.0551
Spain	0.0083	0.0030	0.0036	0.0175

Figure 1 reports the marginal intensities representing instantaneous probabilities of default at sovereign level and Table 1 the corresponding descriptive statistics. The economic impact of the outbreaks and its persistence are evident from the results. Looking at the whole time window Jan 2016/Jan 2021, Southern EU Countries show a systematic higher intensity level, if compared to Northern EU Countries. Table 1 highlights the presence of differences in the order of magnitude for specific descriptive statistics on the intensity levels $\hat{\lambda}_{C_{jt}}$, reflecting the corresponding difference in the embedded riskiness[2]: $\hat{\lambda}_{C_{jt}} \in [0.0008, 0.0084]$ for Northern EU Countries, while $\hat{\lambda}_{C_{jt}} \in [0.0036, 0.2379]$ for Southern EU Countries. Regarding Northern EU Countries, UK is ranked first (e.g. riskiest) over most of the time window[3]. The ranking among these intensities is: UK, France, Netherlands and Germany. By directly comparing the average intensity levels on Dec-2019 and Mar-2020, empirical evidence shows increases driven by scaling factors ranging from 1.4 (Netherlands) to 2.48 (Portugal). Regarding Southern EU Countries, results show Italy and Greece even more affected than Portugal and Spain since the start of the pandemic in 2019. To further investigate the impacts, focusing on Q1 2020 and considering relative changes in the intensity level enable a comprehensive comparison around the beginning of the pandemic event outbreak. When computing relative changes in the intensity levels in Q1 2020 across EU Countries, results show a peak for all countries, with relative increases[4] lying in the interval [38 %, 96 %]. Figure 2 and Table 2 suggest an interesting economic evidence from the pair-wise rank correlation estimates at country level in terms of co-movements. For Northern EU Countries, while Germany-France (France-Netherlands) correlation has the highest (lowest) values until Q3 2020 (Q4 2020), data points are located

[2] Observe also the column “Mean”, reporting the average intensity levels over the period.

[3] Exception for Q1 2017 where France shows a higher peak. From an economic perspective, disentangling the impacts deriving from Brexit would be an interesting topic to analyze for UK: this is beyond the scope of the present paper and left as extension for future research.

[4] The relative changes in the intensity levels are computed by considering the absolute variation in $\hat{\lambda}_{C_{jt}}$ (variation over one time step) divided by the initial intensity level $\hat{\lambda}_{C_{jt}}$. The highest relative changes are reported for Portugal, Spain, France and Greece, within the interval [67 %, 96 %].

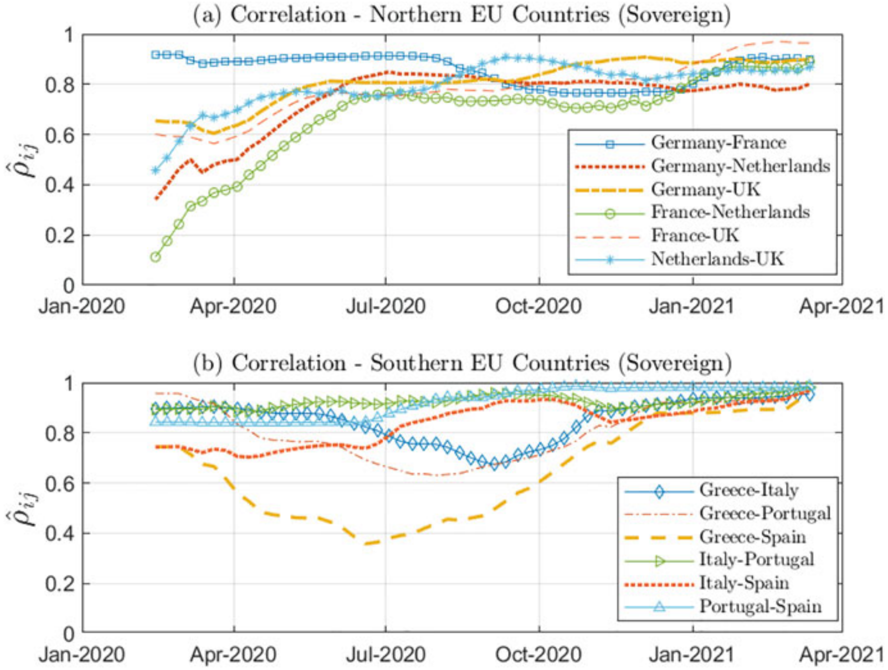


Fig. 2 Sovereign rank correlation. The plots report the pair-wise rank correlation computed based on CDS quotes for 8 EU Sovereign entities over the period Mar-2020/Mar-2021. The plot at the top provides the pair-wise rank correlations for 4 Northern EU Countries, e.g. UK, Germany, France, Netherlands. The plot at the bottom provides the pair-wise rank correlations for 4 Southern EU Countries, e.g. Greece, Italy, Portugal, Spain

Table 2 Descriptive statistics. Rank correlation. Time window: Mar-2020/Mar-2021

Northern EU Sovereign	*Mean*	*Std dev*	*Min*	*Max*
Ger-Fra	0.8583	0.0587	0.7638	0.9192
Ger-Ned	0.7375	0.1324	0.3408	0.8493
Ger-UK	0.8108	0.0900	0.6045	0.9085
Fra-Ned	0.6737	0.1848	0.1118	0.8902
Fra-UK	0.7791	0.1111	0.5634	0.9694
Ned-UK	0.7955	0.0938	0.4569	0.9065
Southern EU Sovereign	*Mean*	*Std dev*	*Min*	*Max*
Gre-Ita	0.8522	0.0835	0.6742	0.9542
Gre-Por	0.8069	0.1138	0.6305	0.9680
Gre-Spa	0.6437	0.1910	0.3583	0.9441
Ita-Por	0.9249	0.0240	0.8845	0.9817
Ita-Spa	0.8374	0.0829	0.7034	0.9666
Por-Spa	0.9219	0.0629	0.8409	0.9861

Table 3 Descriptive statistics. EU common factor estimates

Sovereign	Mean	Std dev	Min	Max
Northern EU	0.0128	0.0035	0.0038	0.0184
Southern EU	0.1222	0.0425	0.0588	0.2625

above a correlation value of 0.6 from Q2 2020 onward for all cases. Overall, the rank correlation has an increasing trend until Q2 2020 for all countries and then a more stable behaviour around a high correlation value. For Southern EU Countries, the same initial increasing trend before Q2 2020 is not a common feature. This is strengthened after Q3 2020 and reaches its maximum in Q1 2021, when all series are above 0.9, thus reflecting the interconnectedness of the economies, the associated riskiness and strong co-movement behaviour[5]. The estimates of the common factor in Table 3 highlight a difference of one order of magnitude between Northern EU and Southern EU Countries. The estimates would need further investigation to disentangle the underlying components, both from an economic and mathematical perspectives. A Marshall-Olkin copula model [5] could be used to further explore the interplay between the sovereign default risk and the dynamics of the banking sector, as in [6], together with non-systematic components. Moreover, a comparison with alternative benchmark models for the dependence structure would be desirable. This is beyond the scope of this first empirical investigation and is left for future research on systemic risk attribution.

4 Conclusion

Building on the work by [1], this paper focuses on a sovereign CDS analysis of systemic risk assuming a macroprudential perspective. The paper proposes a data-driven approach applied to CDS quotes to estimate a reduced form model for the marginal intensity of defaults at country level. It considers 8 European sovereign obligors to assess the pandemic implications on the economies. According to results, a systematic effect in default intensities, rank correlation and common factor is observed in the sample. However, geographical differences in terms of macroeconomic perspective (e.g. rank correlation) and credit risk perspective (e.g. marginal intensity, common factor) are present. Empirical evidence highlights a strong co-movement on pair-wise correlations between sovereign default intensities. These preliminary results suggest the basis for a deeper analysis to identify the factors driving heterogeneity across countries. Future research should focus on developing a comprehensive mathematical framework to simultaneously asses: i) the interplay between sovereign-banking system defaults and non-systematic components, ii) the interconnectedness of the banking system, by means of complex

[5] Rank-correlations associated to Greece have a U-shaped behaviour, with a minimum for Q2-Q3 2020. The interplay between financial and non-financial effects could be a relevant driver.

theory and specific dependence metrics (as alternative to [5]). From a macroprudential perspective, designing a formal framework enabling to identify early warning economic indicators would be important to support policy-makers decisions on systemic risk and financial stability.

Acknowledgments The author would like to thank two anonymous referees, the editor and Cees Dicks for comments and feedback.

References

1. A. Baglioni, U. Cherubini, Within and between systemic country risk. Theory and evidence from the sovereign crisis in Europe, Journal of Economic Dynamics & Control, 37 (2013), 1581–1597.
2. BIS (2011), The impact of Sovereign Credit Risk on Bank Funding Conditions, CGFS Papers, 43.
3. S. Cevik, B. Öztürkkal, Contagion of Fear: is the Impact of COVID-19 on Sovereign Risk Really Indiscriminate?, IMF Working Paper, WP/20/263
4. G. Farina, R. Giacometti, M.E. De Giuli, Systemic risk attribution in the EU, Journal of the Operational Research Society, 70(7) (2018), 1115–1128
5. Marshall A.W., Olkin I., A multivariate exponential distribution, Journal of American Statistical Association, 62(317) (1967), 30–44.
6. R. Giacometti, G. Torri, G. Farina, M.E. De Giuli, Risk attribution and interconnectedness in the EU via CDS data, Computational Management Science, 17 (2020), 549–567.

Modelling Ozone Disinfection to Prevent Covid-19 Transmission

Sam Rolland, Hamid Tamaddon Jahromi, Jason Jones, Alberto Coccarelli, Igor Sazonov, Chris Kershaw, Chedly Tizaoui, Peter Holliman, David Worsley, Hywel Thomas, and Perumal Nithiarasu

Abstract A modelling approach is proposed to study ozone distribution and destruction in indoor spaces. The level of ozone gas concentration in the air, confined within an indoor space during an ozone-based disinfection process, was modelled. The emission and removal of ozone from the air volume were carried out using a generator located in the middle of the room. The computational fluid dynamics (CFD) model proposed accounts for ozone generation and decay kinetics, and buoyancy variations in the airflow. This framework was validated against experimental measurements at different locations in the room during the disinfection cycle. The model was then applied to a more challenging environment and demonstrated the suitability of ozone circulation as a disinfection process. The study also highlights the need for a well-controlled ozone removal process.

1 Introduction

The transmission of Covid-19 and many other pathogen can be disrupted by use of strong oxidisers in the disinfection process, [1]. Ozone, being a gas, has the advantage of reaching aerosol particles as well as surface-deposited pathogens, [2, 3]. As an oxidiser, it also has severe drawbacks for human health if inhaled [4], particularly in people subject to respiratory conditions, [5]. The use of ozone to control the spread of Covid-19 is therefore of high interest but must also be used very carefully to avoid adverse effects, which motivates the present study. The work builds on prior research [6] to extend the model validation and investigate how well a portable ozone generator performs in disinfecting teaching spaces.

S. Rolland (✉) · H. T. Jahromi · J. Jones · A. Coccarelli · I. Sazonov · C. Kershaw · C. Tizaoui · P. Holliman · D. Worsley · H. Thomas · P. Nithiarasu
Swansea University, Swansea, UK
e-mail: s.rolland@swansea.ac.uk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_69

2 Method

The study relies on the formulation of the computational model to reproduce the physics involved in ozone dispersion and its implementation as user routines in a finite volume code (Ansys Fluent). Experiments were run for validation and compared to the numerical results. Finally, a new case was run with a more challenging geometry was run to observe whether single-point ozone generation can provide a satisfactory result where a recess is present in the room.

2.1 Numerical Method

The flow problem is solved numerically using a finite volume formulation based on the incompressible conservation equations of mass (1) and momentum (2):

$$\nabla \mathbf{u} = 0, \tag{1}$$

$$\rho\,(\partial \mathbf{u})\,/\,(\partial t) = (\mu + \mu_t)\,\nabla^2 \mathbf{u} - \rho(\mathbf{u}\nabla)\mathbf{u} - \nabla p + \mathbf{F}. \tag{2}$$

In the equations above, ρ is density, $\mathbf{u}$ is velocity, μ is molecular viscosity, and p is pressure. $\mathbf{F}$ is a buoyancy source term based on the solutal expansion, $\beta_c = 0.001$, and the concentration gradient, $(C - C_\infty)$:

$$\mathbf{F} = g\,\beta_c(C - C_\infty)\,\hat{\mathbf{y}}. \tag{3}$$

The vector $\hat{\mathbf{y}}$ above is the unit vector of direction of application of gravity g. The reference concentration of ozone for the study is taken to be $C_\infty = 0$. The transport of ozone concentration is solved with an additional conservation equation:

$$\rho\,(\partial C)\,/\,(\partial t) = -\rho(\mathbf{u}\,\nabla)C + D\nabla^2 C + \rho\,(S - kC)\,. \tag{4}$$

The value for diffusion coefficient of ozone in air was set as $D = 3.1116\ 10^{-5}$ m^2 s^{-1}. The terms S and (kC) are a volumetric source term and a destruction term respectively. The destruction is a first-order concentration-dependent function characterised by a decay rate k.

The turbulent properties of the flow are resolved using the Spalart-Allmaras formulation, adding one transport equation to be solved for the eddy viscosity $\hat{\nu}$ [7, 8].

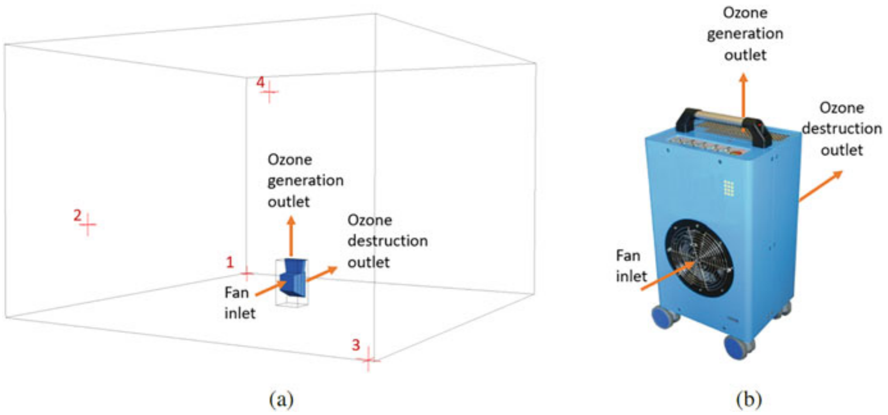
$$\frac{\partial \hat{\nu}}{\partial t} + \nabla\left(\hat{\nu}\mathbf{u}\right) = \frac{1}{\sigma_v}\nabla\left(\left(\frac{\mu}{\rho} + \hat{\nu}\right)\nabla\cdot\hat{\nu} + C_{b2}\left(\nabla\hat{\nu}\right)^2\right) + C_{b1}\hat{S}\hat{\nu} - \left(C_{w1}\hat{\nu}f_w\right). \tag{5}$$

The dynamic turbulent eddy viscosity μ_t in Eq. (2) is calculated using $\hat{\nu}$ and a wall function. For concision, readers are referred to text book materials [9].

2.2 Experimental Work

The experimental validation work was carried out in a simple cuboid room of dimensions 4 m×4.3 m×2.7 m shown in Fig. 1a. A commercially available ozone generator was used to generate and destroy ozone (Duo20, Advanced Ozone Products, UK). The Duo20 unit works in generation mode with a flow rate of 335 m^3 h^{-1}, and a generation rate of 7 g h^{-1}. In ozone destruction mode, the flow rate is 245 m^3 h^{-1}, with an ozone decay rate $k = 0.1$ min^{-1}. The unit can also be used in circulation-only mode, where the flow rate is the same as that used in generation but no ozone is produced (Fig. 1b).

Cycle 1 was run experimentally, and subsequently modelled (Table 1). Two more cycles were run in the simulations for discussion. The ozone distribution was sampled using a BMT 932 ozone monitor in four locations shown in Fig. 1a chosen to sample varied flow conditions: direct flow above the generator (sensor 4), in the corner of the room (sensor 1), near the room ventilation outlet (sensor 3) and against the wall (sensor 2).



(a)

(b)

Fig. 1 Experimental set up of the flow and measurement conditions. (**a**) Room used for the experiment with sensor locations and (**b**) flow through the ozone generator in generation and destruction modes

Table 1 Timings of the ozone disinfection cycles

	O^3 generation	Circulation	O^3 destruction
Cycle 1	180 s	120 s	600 s
Cycle 2	240 s	60 s	600 s
Cycle 3	300 s	0 s	600 s

3 Results and Discussion

3.1 Validation Case

The CFD predicted circulation of ozone is shown in Fig. 2. The data obtained from the validation cases shows that a reasonable agreement can be obtained between the experimental and the simulated data, however this is done without replicating the exact experimental running conditions (Fig. 3). It is apparent that the ozone concentration continues to increase during what is intended as the circulation-only phase of the cycle. This was initially interpreted as an homogenisation of the ozone concentration in the room, but soon found insufficient to explain the continued rise [6]. Two more cycles were modelled with longer generation cycles. These match the experimental data more closely. It is the authors' interpretation that switching the device to a circulation cycle does not end the ozone generation immediately, and some hysteresis exists in the shut-down of ozone generation. As soon as the generator is in destruction mode, the model is able to reproduce experimental results, thus showing that generation and destruction rates can be modelled accurately. These observations are valid at all four sensor locations, thus supporting a good numerical reproduction of the ozone circulation experimentally observed.

3.2 Extension Case

The use of the second case enabled an insight into a geometrically more complex laboratory environment with a recess above the worktops, below wall-mounted

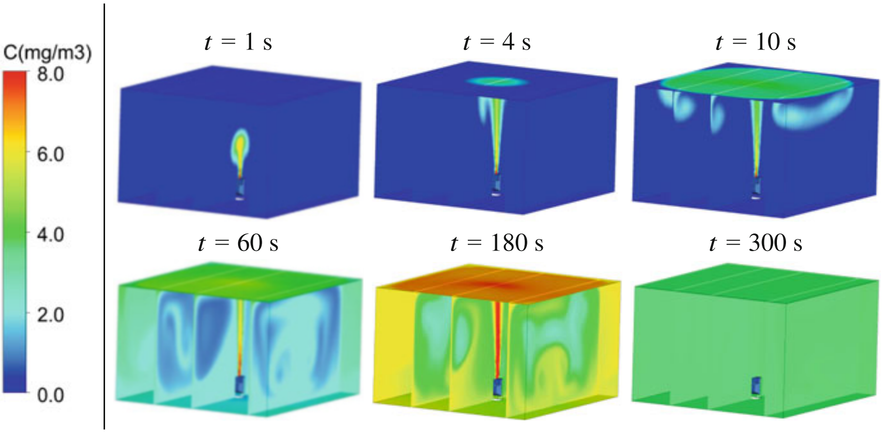
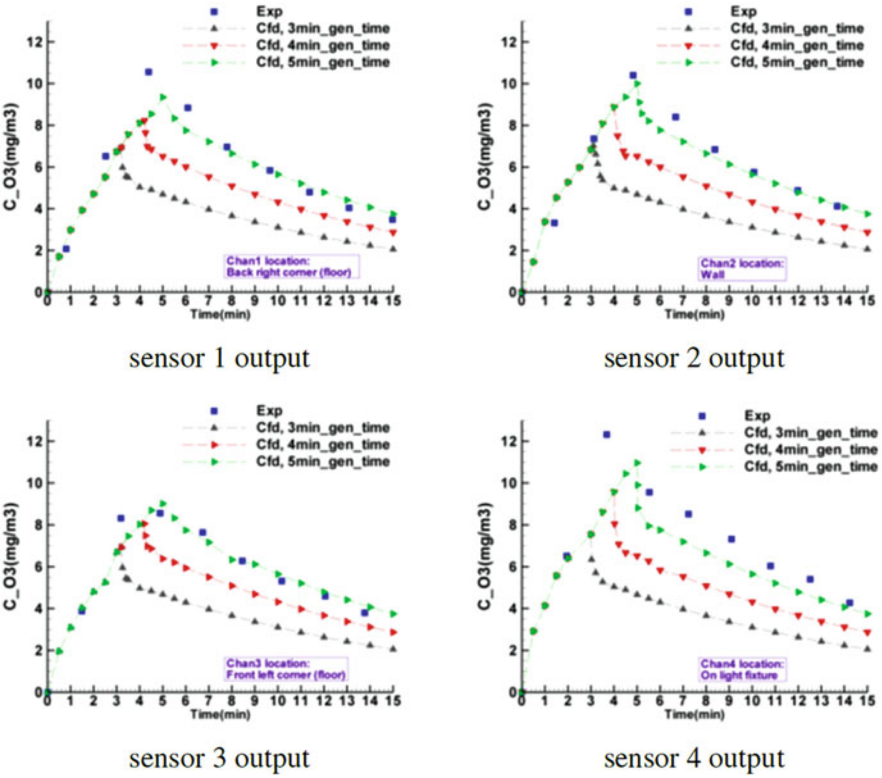


Fig. 2 Ozone concentration in the simple room used for validation



sensor 1 output

sensor 2 output

sensor 3 output

sensor 4 output

Fig. 3 Time-series of ozone concentration at each sensor location (experiment and simulations)

cupboards. These are the areas at risk when using a single-source in a disinfection cycle. The room is 43% larger than the room in the validation case (6 m × 3.6 m × 2.9 m). The density was normalised by the highest ozone density achieved in the simulation to focus on distribution independent of the value assigned to the ozone flux S. The generation-only cycle shown in Fig. 4 represents 300 s. It is notable that within 300 s all relative concentration is above 0.8, including the areas in the recess of concern without a circulation cycle.

4 Conclusion

The method developed was shown to describe the convection and diffusion of the ozone solute phase accurately using a simple turbulence model (Spalart-Allmaras), a buoyancy term formulated to couple phase concentration and convection, and a surface flux and first order volumetric source terms for generation and destruction boundary conditions. Numerical results are in good agreement with the experi-

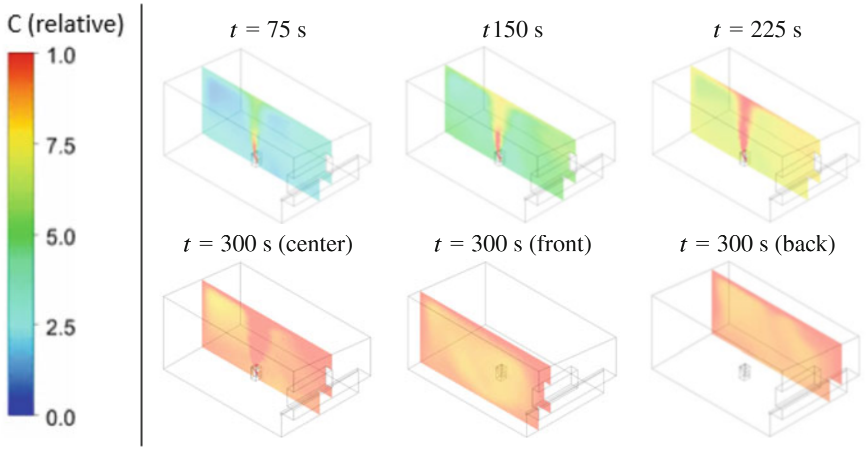


Fig. 4 Relative density contours of ozone concentration in the laboratory

mental measurements. The study showed that issues are not so much in transport modelling as in the faithful reproduction of the generation cycle, principally attributed to experimental uncertainty. The application of the model to a new room geometry shows that diffusion of ozone ensures that a single-source ozone generator also works for a significantly larger room with non-uniform features likely to be areas where the users of the room may come in direct contact with the the virus.

References

1. E. Grignani, A. Mansi, R. Cabella, P. Castellano, A. Tirabasso, R. Sisto, M. Spagnoli, G. Fabrizi, F. Frigerio, and G. Tranfo, *Safe and effective use of ozone as air and surface disinfectant in the conjuncture of COVID-19*, Gases, 1(1):19–32, 2021.
2. C. Tseng and C. Li, *Inactivation of surface viruses by gaseous ozone*, J. Environmental Health, 70(10):56–63, 2008.
3. C. Tizaoui, *Ozone: A Potential Oxidant for COVID-19 Virus (SARS-CoV-2)*, Ozone: Science & Engineering, 42(5):378–385, 2020.
4. D.B. Menzel, *Ozone: An overview of its toxicity in man and animals*, J. Toxicology Environ. Health 13(2-3):181–204, 1984.
5. N.A. Molfino, S.C. Wright, I. Katz, S. Tarlo, F. Silverman, P.A. McClean, A.S. Slutsky, N. Zamel, J.P. Szalai, and M. Raizenne. *Effect of low concentrations of ozone on inhaled allergen responses in asthmatic subjects*, The Lancet, 338(8761):199–203, 1991. Originally published as Volume 2, Issue 8761.
6. H.R.T. Jahromi, S. Rolland, J. Jones, A. Coccarelli, I. Sazonov, C. Kershaw, C. Tizaoui, P. Holliman, D. Worsley, H. Thomas, et al, *Modelling ozone disinfection process for creating COVID-19 secure spaces*, Int. J. Numer. Meth. Heat & Fluid Flow, 2021.
7. P. Spalart and S. Allmaras, *A one-equation turbulence model for aerodynamic flows*, In: 30th Aerospace Sciences Meeting and Exhibit, page 439, 1992.

8. S.R. Allmaras and F.T. Johnson, *Modifications and clarifications for the implementation of the spalart-allmaras turbulence model*, In: Seventh International Conference on Computational Fluid Dynamics (ICCFD7), pages 1–11, 2012.
9. H.K. Versteeg and W. Malalasekera, *An introduction to computational fluid dynamics: the finite volume method*, Pearson education, 2007.